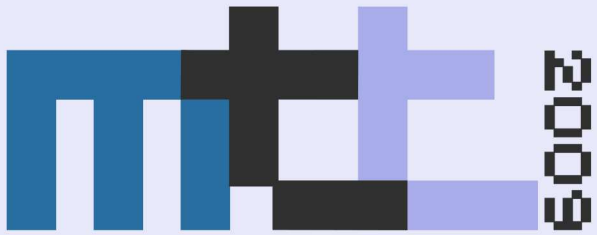


meaning text theory

Proceedings

Fourth International Conference on Meaning-Text Theory

Quatrième conférence internationale sur la Théorie Sens-Texte

Actes

editors: David Beck, Kim Gerdes, Jasmina Milićević and Alain Polguère

ISBN 978-2-9811149-0-7

MTT 2009

Local organization chair

Alain Polguère, Université de Montréal

Program chairs

David Beck, University of Alberta

Jasmina Milićević, Dalhousie University

Program committee

Margarita Alonso Ramos, University of Coruña
Valentina Apresjan, Russian Academy of Sciences
Lorraine Baqué, Autonomous University of Barcelona
David Beck, University of Alberta
Igor Boguslavskij, Russian Academy of Sciences/Polytechnical University of Madrid
Gaétane Dostie, Université de Sherbrooke
Thierry Fontenelle, Microsoft Natural Language Group
Bernard Fradin, Université Paris 7
Kim Gerdes, Université Paris 3
Marie-Josée Hamel, Dalhousie University
Eva Hajičová, Charles University
Leonid Iomdin, Russian Academy of Sciences
Lidija Iordanskaja, Université de Montréal
Sylvain Kahane, Université Paris 10
Richard Kittredge, CoGenTex Inc.
Irina Levontina, Russian Academy of Sciences
Marie-Claude L'Homme, Université de Montréal
Igor Mel'čuk, Université de Montréal
Jasmina Milićević, Dalhousie University
Alain Polguère, Université de Montréal
Owen Rambow, Columbia University
Tilmann Reuther, University of Klagenfurt
Agnès Tutin, University Stendhal-Grenoble 3
Serge Verlinde, Université catholique de Louvain
Leo Wanner, University Pompeu Fabra
Daniel Weiss, University of Zurich

Foreword

We are pleased to present the proceedings of the 4th International Conference on Meaning-Text Theory (MTT '09), held from June 16 to 18, 2009, at the Université de Montréal. The papers contained in this volume bring together the work of 39 researchers from around the world, and are representative of the growing number of people and the diversity of interests within the Meaning-Text research community. This year's proceedings include work from a number of younger scholars, as well as papers from more established researchers with solid international reputations.

These proceedings include papers on subjects from a wide number of areas including theoretical linguistics, translation, computational linguistics, natural language processing, and applied linguistics, focusing on a variety of languages, ranging from familiar Indo-European languages to Mandarin Chinese, Wolof, and Dene Sųłiné. In order to make the papers available to the wider research community, these proceedings are being published electronically and distributed freely at http://www.meaningtext.net.

We extend our heartfelt thanks to all the participants for their contribution to the event, and offer special thanks to our guest speakers, Juri Apresjan and Robert Van Valin Jr., as well as to the Program Committee for their careful work on paper selection. Last but not least, we would like to express our appreciation to Emmanuel Chieze, Patrick Drouin, Marie-Claude L'Homme, and Benoit Robichaud of the Observatoire de linguistique Sens-Texte (OLST) of the Université de Montréal for their invaluable help with local organization. Financial support for MTT '09 was generously provided by the Fond Québecois de Recherche sur la Société et la Culture (FQRSC) and the Department of Linguistics and Translation of the Université de Montréal

David Beck, Kim Gerdes, Jasmina Milićević and Alain Polguère

Index

The Theory of Lexical Functions: An Update[1]

Juri Apresjan

Institute for Information Transmission Problems of RAS
127994, Moscow,
Bolshoi Karetny 19
apr@iitp.ru

Abstract

The standard theory of lexical functions (LFs) roughly claims: (a) that the values of simple LFs of the OPER-LABOR-FUNC family are semantically void; (b) that the values of collocate LFs are phraseologically bound with regard to their argument words and that, consequently, collocations of the form *L* + *X*, where *L* is the value of a certain LF from the argument *X*, are idiomatic intra- and interlinguistically. The present paper purports to show that all the lexemes which are values of collocate LFs are meaningful and that their choice is semantically quite well, though not completely, motivated. The basic assumption for both claims is that semantically well-formed sentences are subject to the general law of semantic agreement which requires of collocated items *L* and *X*, with the exception of a small number of genuinely idiomatic combinations, one of the following two things: either *L* or *X* meets the semantic conditions for filling the valency of the other; or both collocated items display at least one recurrent semantic component in their lexical meaings. These considerations create a foundation for at least partially predicting a set of probable values for each LFi(*X*). Such plausible lexicographic expectations allow to proceed from an item-by-item description of lexicon to a description of the vocabulary of a language as a system.

1 The Standard Theory of LFs

In the last quarter of the 20-th century theoretical linguistics witnessed a breakthrough of paramount importance in the study of constrained lexical co-occurrence. I mean the theory of lexical functions (LFs) proposed by Igor Mel'čuk and Alexander Zholkovsky; references are too well known to be necessary.

I shall confine myself to a discussion of only one portion of the theory bearing on the collocate LFs, otherwise called parameters.[2] Even out of the set of collocate LFs I shall take up a very small selection. My principal concern will be the following LFs: **MAGN**, **OPER1**, **OPER2**, **LABOR1-2, ADV1**, and **ADV2**; occasionally I shall mention some other LFs.

Two main properties were ascribed to collocate LFs in the early work of the two authors.

a) The simple LFs of the **OPER-FUNC** family are semantically void. It follows from the fact that such sets of sentences as, for instance,

(1) a. *The government controls* [X] *all foreign trade*
b. *The government has <exercises>* [OPER1(S0(X))] *control* [S0(X)] *over all foreign trade*
c. *All foreign trade is under* [OPER2(S0(X))] *the control* [S0(X)] *of the government*

[1] This research has been supported in part with grants from the Russian Federation Government (No. НШ-3205.2008.6), the Russian foundation for fundamental research (No. 08-06-00344), and the History and philology department of RAS within the Program "Genesis and interaction of social, cultural, and language communities".

[2] I prefer to speak of "collocates" rather than "parameters" because the latter term has come to be used as a label for a certain semantic class of nouns, such as *height, length, speed, pressure* and the like.

are paraphrases of one another, i.e. synonymous with regard to the situation described. Since the collocations *to have control over something* and *to be under somebody's control* are referentially identical to the core verb *to control* in (1a), it seems natural to conclude that all of the lexical meaning of the collocation is contained in the argument noun *control*, while *to exercise* and *to be under* serve exclusively the auxiliary function of "verbalizing" the noun.[3] In other words, their semantic contribution to the meaning of the utterance is null; see, for example, (Mel'čuk, 1974: 93) and (Mel'čuk et al., 1999: 77).

b) The values of collocate LFs are phraseologically bound, i.e. idiomatic (Mel'čuk, 1974: 80). In other words, the choice of lexeme *L* as value of a certain LF from argument *X* is semantically unmotivated, and the whole collocation *L + X* is idiomatic intralinguistically and interlinguistically. A handy example is LF **MAGN**. We usually say

(2) a. *wolfish appetite*, *raging thirst*, rather than
b. *?raging appetite,* or *?wolfish thirst*

(3) a. *deadly asleep, wide awake*, but not
b. **wide asleep* or **deadly awake*

(4) a. *inveterate liar, inveterate gambler, inveterate smoker*
b. *heavy gambler, heavy smoker*, but hardly
c. **heavy liar*

Interestingly, in Russian the literal equivalents of (4b) are ruled out, and (4c) is equally impossible:

(5) a. **tjazhelyj igrok* 'heavy gambler', * *tjazhelyj kuril'ščik* 'heavy smoker'
b. * *tjazhelyj lžec* 'heavy liar'

The right way to put these ideas in Russian is

(6) a. *zajadlyj <zavzjatyj> igrok* 'inveterate gambler'
b. *zajadlyj <neispravimyj> kuril'ščik* 'heavy smoker'
c. *neispravimyj <ot''javlennyj> lžec* 'inveterate liar'

At first glance at these examples the choice of the adjective to express the meaning of 'high degree of what is denoted by the keyword' seems unaccountable. Hence the conclusion that such collocations are idiomatic intralinguistically and interlinguistically.

As a matter of fact both these assertions were slightly qualified even in the first versions of the theory. In (Mel'čuk & Zholkovsky, 1984: 54) the LFs of the **OPER-LABOR-FUNC** family were defined as verbs "which are semantically empty in the context of the entry lexeme (= their keyword)", with the following continuation in a footnote: "These verbs may be genuinely empty, as, for instance the verb *okazat'* (which does not mean anything definite and cannot be translated into a different language out of context), or meaningful, but then their own meaning is included in the sense of their keyword" (ibid., 96). See also (Mel'čuk, 1974: 104), (Mel'čuk et al., 1992: 32), (Mel'čuk & Wanner, 1996), with the same assertion. As concerns idiomaticity, Mel'čuk (1974: 105) notes that LF collocations on the whole are less idiomatic than genuine idioms, although some of them are very close to the latter.

Similar ideas were voiced in later work; see, for instance, (Uspensky, 1979), (Paducheva, 1991) and (Reuther, 1994, 1996, 2003). However, all of the quoted sources, except Uspensky (1979), relied mostly

[3] The collocation *to be under control* whose lexical meaning may seem to be different from that of *to control* is semantically and syntactically equivalent to the passive form of the verb: *All foreign trade is under the control of the government = All foreign trade is controlled by the government.* Since the passive form of a verb is assumed to possess the same lexical meaning as its active form, the same should apply to the collocation *to be under control* and the core verb *to control*.

on the material of **OPER1**, especially on the argument lexemes denoting emotions, in which case **OPER1** is more or less uniformly expressed by the verbs like *to feel* (*admiration, contempt, envy, fear, indignation, jealousy, pride, shame*, etc.) and therefore is readily interpreted as semantically motivated.

2 An Update: General Considerations

In connection with my practical lexicographic research I have also taken up the subject. I have followed up the general trend of reasoning outlined in the previous work, but have made a particular point of extending it in several directions. In this section I shall give a brief sketch of the update proposed in (Apresjan, 2004, 2008a, 2008b), (Apresjan & Glovinskaja 2007). I shall argue that: 1) all collocate LFs, including the simple LFs of the **OPER-LABOR-FUNC** family, have a lexical meaning of their own; 2) the choice of a particular lexical item *L* as value of a certain LF from the argument lexeme *X* is conditioned by a) the nature of the LF in question, b) the lexical meaning of *L*, and c) the semantic class and subclass of a Vendlerian classification to which *X* belongs.

2.1 General Foundations

The general foundations of the above claims are twofold – paradigmatic and syntagmatic.

2.1.1 Paradigmatic Foundation

The paradigmatic foundation is a Vendler-like classification of predicates into actions (*to look, to read, to walk, to write*), activities (*to trade, to negotiate, to educate, war*), processes (*to flow, to grow, growth*), spatial positions (*to stand, to sit, to lie*), states (*to feel, to see, to know, need*), properties (*authority*, *beauty, courage, to stammer*) and so on, with further subdivisions of these major classes into a number of more compact subclasses like, say, physical states (*to see, to hear*), physiological states (*to itch, to ache*), mental states (*to think, to know*), volitional states (*to wish, to intend*), emotional states (*afraid*, *envy, to fear*), economic states (*need, to prosper*), social states (*married, divorced*) etc.

I should like to emphasize that this classification, if properly modified, is semantically valid with regard to any class of predicates, not only verbs.

On the other hand, it underlies not only the aspectual properties of verbs, as was currently believed, but other verbal grammatical categories like mood and voice as well. A familiar example are mental statives like *to know* which defy the use in the imperative mood and in the passive voice. We usually say

(7) a. *You **should** know how to do it* rather than
b. ??***Know** know how to do it!*

(8) a. *This fact is known **to** everybody* rather than
b. ?*This fact is known **by** everybody*, with a genuine agentive complement

Last, but not least: the semantic classes and subclasses to which a given predicate belongs condition to a large extent not only its purely grammatical properties, but also its government pattern and much of its combinatorial profile.

For example, the overwhelming majority of many place verbal predicates starting from four-actant verbs and on denote actions (not activities, processes, states, or anything else). Notoriously abundant in many-place predicates are semantic domains of causing locomotion, creating physical objects, exchange of valuables between two persons and some others. Here are some examples:

(9) a. *The allies* [A1] *transported troops* [A2] *from Britain* [A4] *to France* [A3] *by aircraft* [A5]

b. *The girl* [A1] *sewed a dress* [A2] *(for her doll) out of leftover pieces* [A3] *on her mother's sewing machine* [A4] *with silk thread* [A5]
c. *NN* [A1] *rented <leased> his house* [A2] *to us* [A3] *for a year* [A5] *for 5000 dollars* [A4]

d. *We* [A1] *rented <leased> this house* [A2] *from NN* [A3] *for a year* [A5] *for 5000 dollars*

As concerns combinatorial profiles, I shall have more to say on the subject a little later.

2.1.2 Syntagmatic Foundation

Syntagmatically the above theoretical claims are based on the well-known fact that the majority of collocations in natural languages are subject to the general law of semantic agreement. This law requires of the collocated items *L* and *X* one or both of the following two things:

(a) they should have at least one non-trivial recurrent (repetitive, common) semantic component in their meanings (or, technically speaking, in their meaning definitions); the greater the number of recurrent components in the phrase or sentence the greater the degree of its semantic cohesion;

(b) a potential actant *Ai* of predicate *P* should meet the semantic requirements for the *i*-th valency of *P*, irrespective of whether this valency is active or passive (for similar ideas see (Iordanskaja & Mel'čuk) in the present volume).

Illustration of the first requirement: in the phrase

(10) *to cook fish and chips*

all the three words are polysemous.

To cook means 1) 'to prepare food for eating by using heat' or (coll.) 2) 'to invent something', as in *to cook a story*.

Fish means 1) 'the flesh of a water animal used for food' or (coll.) 2) 'a person with a salient trait', as in *cool fish, poor fish, odd fish*.

Chip means 1) 'a long thin piece of potato cooked in hot fat or oil' or 2) 'a small piece of silicon used to store and process information in computers'.[4]

Even if we confine ourselves just to those six senses (as a matter of fact, there are many more) we shall get eight possible paths of reading (10), e.g. 'to invent a person with a salient trait and a long thin piece of potato'. However, (10) is unequivocally understood in just one reading (10'):

(10') 'to cook the flesh of a water animal and long thin pieces of potato in hot fat or oil'

The intuitively obvious choice of the only reading (10') as semantically cohesive is ensured by the fact that the number of recurrent senses for it, namely, 'food', 'heat', 'hot (oil)' etc., is the greatest.

Illustration of the second requirement: let us look at the phrases

(11) a. *auburn hair*
b. *??auburn horse <furniture>*

As is well known, the color adjective *auburn* means 'reddish-brown' and normally applies to human hair; this is the semantic requirement that any noun filling the (passive) valency of *auburn* should meet. From this point of view (11a) accords with the law of semantic agreement while (11b) deviates from it.[5]

[4] Meaning definitions here and elsewhere are highly informal.
[5] For what follows the distinction between cases (a) and (b) is of little importance, and for the most part I shall ignore it.

All of the above is more or less common knowledge. I have recalled it to make the following less trivial assertion: the general law of semantic agreement holds good not only for free word combinations but for the overwhelming majority of LF collocations as well.

2.2 Two Illustrations

2.2.1 The Case of *okazyvat'*

I shall start with the allegedly empty Russian verb *okazyvat'* which is used in Modern Russian only as part of LF collocations[6]. Let us look at the following two phrases where *okazyvat'* is opposed to *imet'* 'to have' in the function of **OPER1** from the same noun.

(12) a. *okazyvat' vlijanie na voennyx* 'to exert influence on the military'
b. *imet' vlijanie sredi voennyx* 'to have influence among the military'

The difference in the values of **OPER1** is obviously due to the fact that *vlijanie* in these two sentences is used in two different senses, i.e. represents two different lexemes: *vlijanie 1* in (12a) denotes a kind of pressure, that is an **action**; *vlijanie 2* in (12b) denotes the ability to affect somebody's actions and decisions without using force or orders, that is a certain **property** of a person.

This fundamental semantic difference is directly mirrored in the respective **synonym** series: the synonyms of *vlijanie 1* are nouns like *vozdejstvie* 'action' and *davlenie* 'pressure', while the synonyms of *vlijanie 2* are nouns like *avtoritet* 'authority' and *ves* 'weight' (in the figurative sense).

Indirectly the semantic opposition 'action vs. property' is very consistenly reflected in a number of non-semantic distinctions between *vlijanie 1* and *vlijanie 2.*

Grammatical distinctions: *vlijanie 1* has the plural form, while *vlijanie 2* has not: *različnye vlijanija, kotorym on podvergalsja vo vremja učeby v Garvarde* 'various influences he was subject to during his studies at Harvard', but not ***različnye vlijanija, kotorye on imel v voennyx krugax** 'various influences he had in military circles'.

Derivational distinctions: *vlijanie 1* is derived from the verb *vlijat* 'to influence'; there is no verbal counterpart for *vlijanie 2*. *Vlijanie 2* has a derived adjective *vlijatel'nyj* 'influential', which is impossible for *vlijanie 1.*

Government patterns. *Vlijanie 1* governs the preposition *na* 'on' (see 12a), while *vlijanie 2* governs prepositions and prepositional groups *sredi* (see 12a), *v srede, v krugax* 'among', 'in the circles of': *On imeet bol'šoe vlijanie v voennyx krugax <v teatral'noj srede>* 'He has much influence in military circles <in the theatrical milieu>'.

Combinatorial profiles. *Vlijanie 1* has such adjectival LFs as **BON** and **ANTIBON**: *xorošee <položitel'noe, plodotvornoe> vlijanie* 'good <positive, fruitful> influence' vs. *ploxoe <durnoe, otricatel'noe, pagubnoe, tletvornoe> vlijanie* 'bad <harmful, negative, pernicious, baneful> influence'. None of these adjectives are possible for *vlijanie 2.* On the other hand, *vlijanie 2* has such verbal LFs as **INCEPOPER1** *priobretat' (vlijanie)* 'to acquire influence' and **FINOPER1** *terjat' (vlijanie)* 'to lose influence'. Neither is possible for *vlijanie 1.*

In view of these consistent and persistent distinctions between the meanings of action and property of the word *vlijanie* it seems natural to expect that the values of **OPER1** for *vlijanie 1* and *vlijanie 2* should also reflect this fundamental semantic opposition.

Indeed, people *have* properties, so the property lexeme *vlijanie 2* legitimately co-occurs with the verb *imet'* 'to have' as its **OPER1**. It would also be instructive to look at the other possible value of **OPER1** from *vlijanie 2*, which is the verb *pol'zovat'sja* 'to use'.

---

[6] In the XIX-th century it had the now obsolete meaning of 'to show' which is still preserved in some Russian dialects: *Zarja okažet poljakam kak ničtožen otrjad tvoj* 'The dawn will show to the Polish how small your detachment is' (A. Marlinsky); *Zoloto staralis' ne okazyvat'* (P. Bažov) 'They tried to conceal [= not to show to anyone] the gold they had mined' (both quotes are from the Comprehensive Academic Dictionary of Russian).

(13) a. *On pol'zuetsja bol'šim vlijaniem sredi voennyx* 'He has much influence among the military'

Pol'zovat'sja is definitely actional in its principal meaning (*On vsegda pol'zuetsja nožom, čtoby otkryt' banku* 'He always uses a knife to open a can'), yet in the meaning at issue it is purely stative. This is borne out by the fact that in the function of OPER1(*vlijanie 2*) it has no perfective form:

(13) b. **On vospol'zovalsja bol'šim vlijaniem sredi voennyx*

In the actional meanings of *pol'zovat'sja* the perfective form is quite normal: *pol'zovat'sja <vospol'zovat'sja> nožom* 'to use a knife', *pol'zovat'sja <vospol'zovat'sja> slučaem* 'to avail oneself of the chance'.

Let us now turn to the lexeme *vlijanie 1*. It denotes an action, and actions are *performed* or *done*. It is natural to assume therefore that *okazyvat'* as OPER1(*vlijanie 1*) is "synonymous" to the above verbs, i.e. has the meaning of doing.

This assumption is corroborated with the following fact. There are about thirty collocations of nouns with the verb *okazyvat'* in modern Russian, and in most of them the nouns denote actions:

(14) a. *Okazyvat' vlijanie* 'influence', *vozdejstvie* 'action, impact', *davlenie* 'pressure', *dejstvie* 'action', *milost'* 'a favour', *nažim* 'pressure', *podderžku* 'support', *pomošč'* 'assistance', *soprotivlenie* 'resistance', *uslugu* 'service'

Most remarkably, in some of those collocations, especially with the nouns *vozdejstvie* 'action, impact' and *dejstvie* 'action', *okazyvat'* is, or until very recently used to be, interchangeable with an undoubtedly meaningful and actional verb *proizvodit'* 'to produce'. This is convincingly attested by the Russian National corpus. Two examples will suffice:

(14) b. *Fioletovyj cvet proizvodit ugnetajuščee dejstvie na nervnuju sistemu* 'The violet color produces an oppressing action <effect> on the nervous system'

c. *Buduči nematerial'nym, prostranstvo ne možet proizvodit' vozdejstvie na tela* 'Being non-material, space cannot produce any action on corporeal bodies'

In view of such facts, to postulate a null meaning for *okazyvat'* will amount to saying that a meaningless item may be synonymous to a meaningful one, let alone the fact that it will also mean postulating an inexplicable exception from the law of semantic agreement. So we are forced to the conclusion that *okazyvat'* has a meaning and that it is a very general meaning of 'doing'.

To produce more evidence for my principal claim I shall pursue the same example a little farther. Let us look at another sufficiently large class of nouns collocating with *okazyvat'* and denoting mental or emotional attitudes towards somebody or something.

(15) *Okazyvat' vnimanie* 'attention', *znaki vnimanija* 'signs of respect', *doverie* 'confidence', *počesti* 'honors', *uvaženie* 'respect', *čest'* 'honor'

At first sight this seems to contradict the claim that *okazyvat'* means something like 'to do, to perform, to produce': one cannot *do <perform, produce> attention* or *confidence*. However, if we look closer at the collocations under (15), we shall discover that *okazyvat'* there represents an LF different from **OPER1** and, consequently, cannot mean 'doing'. In the Comprehensive Academic Dictionary of Russian (the only one to single out this meaning) *okazyvat'* in these collocations is defined as 'to display or to show one's attitude to somebody or something', which should be interpreted in terms of LFs as the value of **MANIF**. It is noteworthy that *okazyvat'* in this case is interchangeable with a more or less standard expression of **MANIF** by means of the verb *projavljat'* 'to display, to show':

(16) a. *okazyvat' <projavljat'> doverie (k) komu-libo* 'to display <to show> confidence in somebody'
b. *okazyvat' <projavljat'> uvaženie (k) komu-libo* 'to display <to show> respect for somebody'[7]

Now, if *okazyvat'* in (15) is the value of **MANIF**, what then is the value of **OPER1** for the nouns in question? The standard value of **OPER1** from mental and emotional attitudes is the verb *ispytyvat'* 'to feel, to experience'.

(17) a. *ispytyvat' doverie k komu-libo* 'to feel confidence in somebody'
b. *ispytyvat' uvaženie k komu-libo* 'to feel respect for somebody'

Mental and emotional attitudes semantically neighbor on emotional states like *admiration, fear, indignation, joy, shame* etc., which are *felt* by people.[8] It is no wonder therefore that mental and emotional attitudes share the same **OPER1** with emotional states.

(17) c. *ispytyvat' vosxiščenie <vozmuščenie>* 'to feel admiration <indignation>'
d. *ispytyvat' strax <styd>* 'to feel fear <shame>'

On the other hand, mental and emotional attitudes semantically neighbor on properties, and properties, as has already been pointed out, collocate with the verb *imet'* 'to have' as value of **OPER1**. Interestingly, in Russian and, more commonly, in English, the have-verbs may replace the verbs like *ispytyvat'* for this whole class of arguments. In other words, the lexemes *ispytyvat'* and *imet'* are synonymous in the context of nouns denoting mental and emotional attitudes:

(18) a. *imet' doverie k komu-libo* 'to have confidence in somebody'
b. *imet' uvaženie k komu-libo* 'to have respect for somebody'

We shall see more of the semantic opposition 'emotional attitudes' vs. 'emotional states' below.

Before I proceed to my next example I should like to call attention to the following fact. As is clear from my glosses all the way through, the English lexemes *influence 1* and *influence 2* display almost the same kind of semantic, grammatical, derivational, syntactic, and combinatorial distinctions as their Russian counterparts *vlijanie 1* and *vlijanie 2*. This is the first piece of evidence showing that the degree of interlinguistic idiomaticity of LFs in the standard theory has been somewhat exaggerated.

2.2.2 ADV1, ADV2 and their Compounds

My second illustration will be the LFs **ADV1** and **ADV2**, with various compositions. This time I shall make a special point of emphasizing, apart from other things, that the combinations of the form *L + X* are less idiomatic, than has been believed hitherto, not only interlinguistically but intralinguistically as well.

In English **ADV1** from nouns denoting meals is uniformly expressed by the preposition *at* + the argument lexeme: *at breakfast, at dinner, at lunch, at meal* (*The whole family meets at meals*), *at supper, at table* (= 'while eating'), *at tea.* Their Russian counterparts are uniformly expressed by the combination of the preposition *za* + the argument lexeme: *za zavtrakom*, *za obedom, za lančem, za edoj, za užinom, za*

[7] The preposition *k* is bracketed because *okazyvat'* has a government pattern of its own with a prepositionless dative while *projavljat'* borrows the government pattern of the argument word.
[8] One of the first attempts to explore linguistic differences between the names of emotional states and emotional attitudes was undertaken in (Iordanskaja, 1970).

stolom, za čaem. There is also a uniformly expressed **ADV2REAL1** for such arguments: *for breakfast, for dinner, for supper* etc.; their Russian analogues are *na zavtrak, na obed, na užin* etc.

ADV2 from nouns denoting guidance and supervision is uniformly expressed by the preposition *under* + the argument lexeme: *under somebody's control <direction, guidance, leadership, observation, oversight, superintendence, supervision>*. They also allow of literal translations into Russian: *pod č'im-libo kontrolem <rukovodstvom, voditel'stvom, nabljudeniem, prismotrom, upravleniem, nadzorom>*.

ADV1 from the names of emotional states is uniformly expressed by the preposition *in* + the argument lexeme: *in admiration, in agitation, in amazement, in anger, in anxiety, in bewilderment, in confusion, in delight, in despair, in doubt, in embarrassment, in fury, in horror, in indignation, in panic, in rapture, in sorrow, in surprise, in suspense* etc. They also allow of literal translations into Russian: *v bespokojstve, v gneve, v izumlenii, v jarosti, v otčajanii, v panike, v pečali, v smjatenii, v somnenii, v trevoge, v udivlenii, v užase, v vostorge, v vosxiščenii, v zamešatel'stve*, etc.

The same class of arguments gives rise to one more adverbial LF – **ADV1MANIF** – which can be roughly defined as 'displaying X while doing something'. This LF is expressed by the preposition *with* + the argument lexeme:

(19) a. ADV1(*horror*) = *in horror*, as in *to look at somebody in horror*
b. ADV1MANIF(*horror*) = *with horror*, as in *to look at somebody with horror*

Note also collocations like *with admiration, with anger, with anxiety, with envy, with fury, with incredulity, with indignation, with joy, with pride, with shame, with surprise, etc.,* which also fall under (19b). Their Russian counterparts are uniformly expressed by the combination of the respective preposition *s* + the argument lexeme in the instrumental case: *s bespokojstvom, s gordost'ju, s izumleniem, s jarost'ju, s nedoveriem, s radost'ju, so stydom, s udivleniem, s užasom, s vostorgom, s vozmuščeniem, s zavist'ju,* etc.

There is a semantically motivated difference between the classes of arguments for those two LFs. **ADV1MANIF** is possible from the names of emotional states like *anger, despair, horror, surprise* and so on, as well as from the names of emotional and mental attitudes like *contempt, hatred, interest, love, respect, sympathy* etc., as in *with contempt, with hatred, with interest, with love, with respect, with sympathy.* **ADV1** of the form *in + X* for emotional and mental attitudes is impossible.

The difference is not accidental. Both, states and attitudes can be expressed outwardly; hence the collocations of the type *with admiration* and *with love*. However, it is only states that a person can ***be in***. One cannot *be in an **attitude***.

It is really amazing how extraordinarily sensitive to the minutest semantic distinctions a natural language can be. I mean the distinction between emotional states and emotional attitudes. It also shows up in the next group of examples illustrating the LF **ADV2CAUS(X)** whose meaning can be formulated as 'X having been caused by the current situation P'.[9]

In English **ADV2CAUS** is uniformly expressed by the collocation *to* + the argument lexeme: *to his amazement, to the delight of the crowd, to our discredit*, *to my displeasure, to the horror of the spectators, to his pleasure, to my regret, to his surprise.* In Russian this LF is expressed no less consistently by the collocation *k* + the argument lexeme: *k ego izumleniju, k vostorgu tolpy, k našemu stydu, k moemu neudovol'stviju, k užasu zritelej, k ego udovol'stviju, k moemu sožaleniju, k ego udivleniju.*

This LF is possible for some emotional states and is ruled out for mental and emotional attitudes. Once again, the difference between the two classes of arguments has a profound semantic motivation. Emotions like *amazement, delight, displeasure, horror, regret* and so on are transient, often short-lived inner states which may quickly pass over after the factor causing them has ceased to act on the Experiencer.

[9] In the entries of the Russian TKS written by L. Iordanskaja, where this LF was first singled out, it was identified as **ADV2B**.

Contempt, hatred, interest, love, respect, sympathy and other emotional and mental attitudes resemble properties in that they do not arise as a transient reaction to the current situation but are rather constant.[10]

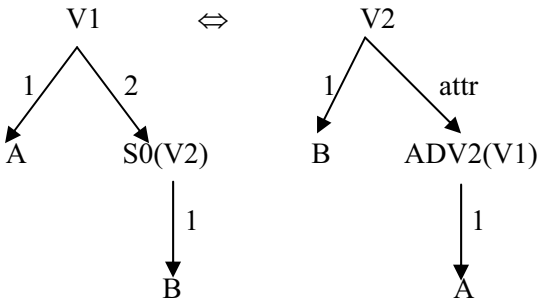
On the basis of these observations one can formulate two new rules of paraphrasing.

(20) a. *A controls the work of B—B works under A's control*
b. *A directs the work of B—B works under A's direction*
c. *A oversees the studies of B—B studies under the oversight of A*
d. *A superintends the studies of B—B studies under the superintendence of A*
e. *A supervises the coaching of B—B coaches under the supervision of A,* etc.

These equivalences can be formalized in the following way (V means a verb):

(20') V1 + S0(V2) ⇔ V2 + ADV2(V1)

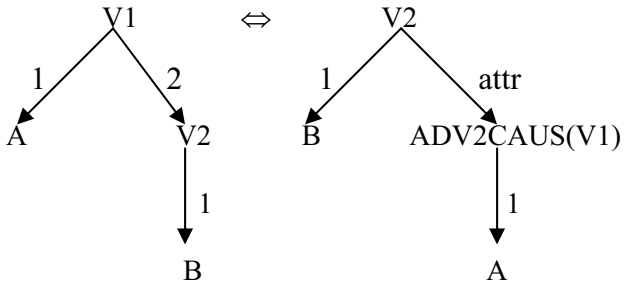
(20")



(21) a. *She was delighted <displeased> that he refused—To her delight <displeasure>, he refused*
b. *Jane was horrified that her son also fell ill—To Jane's horror, her son also fell ill*
c. *I regretted very much that I lost touch with her—To my great regret, I lost touch with her*
d. *I was surprised <astonished> that she came—To my surprise <astonishment>, she came*

(21') V1 + V2 ⇔ V2 + ADV2CAUS(V1)

(21")



3 A Case Study

The starting point of this research was the Russian word *kontrol'* 'control', and the purpose was to account for the choice of *strogij* 'strict' as value of the LF **MAGN** for it. The two respective English words (or, rather, lexemes), *control* and *strict*, are semantically very similar to their Russian counterparts. So to save myself the trouble of glossing the examples I shall go straight to illustrating the point with English material.

[10] More permanent states like *zavist'* 'envy' and *revnost'* 'jealousy' cannot be used in this construction either: **K ego zavisti <revnosti>, ego sopernik polzovalsja ból'šim uspexom* 'To his envy <jealousy>, his rival had a greater success'.

3.1 Semantic Correspondence Between LF Value and Argument Lexeme: *strict control*

Similarly to the related verb, the substantive lexeme *control* is a two-place predicate denoting a hierarchically ordered situation with an Agent having a higher status in the hierarchy and therefore able to influence the actions, states, and sometimes even the very existence of the Patient, no matter if the latter is a person, a body of persons or a certain state of affairs. The idea of dominance of the first actant over the second is also apparent in the meaning of the governed preposition *over* which is by no means "semantically empty":

(22) *The president had firm control over the Cabinet*

Russian uses a very similar preposition *nad* 'over' to express the same syntactic and semantic relation between the two actants of *kontrol'* 'control'.

Let us now turn to the meaning of the adjective *strict* as value of the LF **MAGN** for *control*. *Strict* is a predicate describing similar relations between two participants of a hierarchical situation: the Agent has "the whip hand" over the Patient. Note, for instance, that parents can be *strict* with their children, while the latter can hardly be *strict* with their parents, unless, of course, the respective utterance is intended as a joke. The same holds good for the relations between teacher and pupil, employer and employee, examiner and examinee and so on. In all these cases the participant with a higher biological or social status can be strict with the participant whose respective status is lower, but not the other way round.

In the case at issue semantic agreement assumes the form (a), with the recurrent semantic component being the idea of hierarchical relations between the two participants of the situation: it makes part of the lexical meanings of both lexemes, *control* and *strict*, and thus accounts for the choice of *strict* as MAGN(*control*). Naturally enough, the same is true of Russian which, as has already been mentioned, chooses the adjective *strogij* 'strict' as the value of **MAGN** for *kontrol'* 'control'.

3.2 One Value of LF—a Semantically Homogeneous Class of Arguments

In the preceding section we dealt with one value *L* (*strict*) of one LF (**MAGN**) with regard to one argument lexeme *X* (*control*). However, if the hypothesis of semantic agreement between *L* and *X* in LF collocations is at least partially true, then the same value of the same LF should be possible for a larger class of arguments which are semantically similar to *control.*

Control makes part of a class which comprises such nouns, as, for instance, *audit, censorship, examination, inspection, monitoring, oversight, superintendence, supervision, surveillance*, and, possibly, some other. The above assumption was checked against the material of authoritative British dictionaries (Longman, MacMillan, Oxford Advanced Learners Dictionary) and the retrieval system of the GOOGLE engine, and the search yielded many of the predicted collocations. Here are some examples:[11]

(23) a. *They ordered strict audit of oil and power companies*
b. *Strict censorship (was) imposed on the media in North Korea*
c. *This requires strict temperature control*
d. *They called for a strict examination of export trade refund*
e. *Strict quarantine inspection of US beef products (is considered necessary)*
f. *Their results demand stricter monitoring*
g. *The president promised strict oversight of a $ 787 billion packet he signed into law this week*
h. *Minister Korthals wants stricter supervision of the activities of private detectives*

[11] In case the collocation was not recorded in dictionaries and did not occur at least once in a corpus of a hundred documents recalled by the GOOGLE engine it did not count as existing in the English language.

i. *Strict surveillance combined with quick response could contain bird flu'*

The collocation *strict superintendence* did not occur in any of these sources, but even with this exception the prediction should be considered sufficiently reliable to be used as a lexicographic tool.

3.3 One Argument Lexeme—a Semantically Consistent Class of Values of LFs

If we pursue the same line of reasoning, the next step in it should be the hypothesis that the values *L1, L2, ..., Ln* of all LFs defined for one argument lexeme *X* should feature in their lexical meanings a semantic component which recurs in the lexical meaning of *X*. In this connection it would be natural to look at the principal verbal collocations for *control* which are those of the **OPER-LABOR-FUNC** family. The most conspicuous of them are the ones which involve the second actant of the situation – they demonstrate semantic agreement between the value of a verbal LF and the argument lexeme most convincingly.

(24) a. *to be under <to undergo> (control)* [**OPER2**]
b. *to fall under control* [**INCEPOPER2**]
c. *to put <to place, to bring> (somebody) under (control)* [**CAUSOPER2**]
d. *keep (something) under (control)* [**LABOR1-2**]

The values of these LFs include either the preposition *under*, or the verbal prefix *under-* (in *undergo*). In their physical senses both denote the position **below** something, and enough of this meaning is preserved in the figurative senses like that of *under* (*control*) to ensure semantic agreement between the preposition and the argument lexeme. Since *under* suggests the lower, subjugated position of the second participant, its choice is semantically quite well motivated.

As has been mentioned above, the hierarchical relations inherent in the meaning of *control* can be detected in the values of **OPER1** and some compositions with it as well.

(25) a. *to exercise <to exert> (control) over (something)* [**OPER1**]
b. *to have (control) over (something)* [**OPER1**]

(26) a. *to take <to gain> (control) over (something)* [**INCEPOPER1**]
b. *to lose (control) over (something)* [**FINOPER1**]

In view of the fact that all the LFs in (25) and (26) introduce the first participant whose status is **higher** than that of the second participant, the choice of the preposition *over* with the meaning of 'higher in the hierarchy' seems to be perfectly appropriate on semantic grounds. This conclusion is further substantiated by the relations of antonymy which obtain between the prepositions *over* and *under* representing the relative positions of the first and second actants of the hierarchy respectively.

It is true that *over* in these and similar collocations can sometimes be replaced with the preposition *of* which renders a much more general idea of a **relation** between two entities, not necessarily hierarchical:

(27) a. *to be in (control) of (something)*
b. *to have (control) of (something)*
c. *to take (control) of (something* etc.

It view of such examples I should like to emphasize once again that I do not suggest there is a hundred percent semantic agreement between the lexical meanings of *L* and *X* in an LF collocation *L + X*. The claim I am making is a little more modest: there is a sufficient number of collocations where the choice of verbs as values of certain LFs is so consistent as to leave no doubt that it is semantically motivated.

3.4 A Class of Argument Lexemes—a Semantically Consistent Class of Values of LFs

If there is semantic agreement between LF-values *L1, L2, ..., Ln* and the lexical meaning of *X*, then it is reasonable to expect that the combinatorial profiles of a whole class of arguments which are semantically similar to *X* would be largely coincident.

Before I produce evidence in favor of this assumption I should like to call attention to an interesting semantic peculiarity in the meaning of the noun *control* which makes it partly comparable to the noun *influence* analyzed above. As has been shown in paragraph 2.2.1, the word *influence* breaks up into two distinct lexemes: *influence 1*, which is an action (as in *to exert influence on somebody*) and *influence 2* which a property (as in *to have influence among the military*). We have posited two distinct lexemes for *influence* because the number of differences in grammatical forms, government patterns, combinatorial profiles, synonyms, and derivatives between them is so great as to preclude the possibility of blending them into a single entity.

The noun *control* displays a similar duality of meaning, only in this case it is the opposition of action or activity vs. a certain state of affairs, or, for simplicity's sake, just a state. Consider collocations under (28) in which the actional features of *control* come to the foreground, and collocations under (29) where its stative features are foregrounded.

(28) a. *to exercise <to exert> (control) over (something)* [**OPER1**]
b. *to undergo (control)* [**OPER2**]

(29) a. *to have (control) over (something)* [**OPER1**]
b. *to be under (control)* [**OPER2**]

However, in the semantic structure of *control* the opposition 'actional' vs. 'stative' has not solidified to a degree when the opposed senses constitute two distinct entities: grammatical forms, government patterns, derivatives, and synonyms, if any, are the same, so both uses co-exist within a single lexeme.

With regard to this opposition the lexemes of the semantic class to which *control* belongs fall into three subclasses: 1) the subclass of purely actional lexemes, such as *audit, checkup, (university) examination, inspection, monitoring, tests, trials*; 2) the subclass of purely stative lexemes, such as *care, charge, power*; 3) the subclass of lexemes which allow of both types of uses, actional as well as stative, such as *control, supervision, surveillance*.

Proceeding from our assumptions we can expect to find significant intersections in the combinatorial profiles of those lexemes. In view of the limitations of space I shall have to confine myself to considering just one LF to substantiate this claim. Let it be **OPER2**.

Our lexicographic expectations with regard to the probable values of **OPER2** from these argument lexemes can be formulated as follows: 1) the actional lexemes of the first class will collocate with the verb *to undergo*; 2) the stative lexemes of the second class will collocate with the verbal group *to be under*; 3) the intermediate lexemes of the third class will allow of both these values of **OPER2**.

The data of the American corpus bear out all of these expectations.

(30) *to undergo an audit <a checkup, an examination, an inspection, monitoring, tests, (sea) trials>*

(31) *to be under the care <the charge, the power> (of somebody)*

(32) a. *to undergo control <supervision, surveillance>*
b. *to be under control <supervision, surveillance>*

It is noteworthy that the collocations of the type

(33) a. *??to be under an audit <a checkup, an inspection, monitoring, ...>*
b. **to undergo the care <the charge, the power> of somebody*

have not been corroborated by the corpus.

4 Conclusion

LFs collocations form a continuous space with two poles: (a) highly idiomatic collocations, like **MAGN** *wolfish appetite, raging thirst, wide awake, inveterate liar,* etc., with LF values which are possible only for a very limited number of argument lexemes and are therefore semantically hazy or downright unaccountable; (b) collocations like **BON** *bad behavior* (*improper* is more idiomatic), *bad deal* (*raw* is more idiomatic), *bad effect* (*harmful* is more specific), *hear badly* (*indistinctly* is more specific), *bad influence* (*baneful* is more idiomatic) etc., with LF values which are possible for a very wide range of argument lexemes and are therefore semantically clear and to a large extent predictable. The former border on pure idioms, and the latter border on free word combinations (though they are not quite free).

In between there are collocations which form the bulk of LF material. With regard to such collocations —and they have been our principal concern—one can formulate useful lexicographic expectations as to their LF potential and the possible values of concrete LFs. This allows to proceed from an item-by-item lexicographic description to a much more systematic treatment of material by compact classes of lexemes.

On the other hand it seems to be a plausible claim that the regularities we have observed reflect language competence of the speakers and should therefore be taken into account in constructing meaning-text models for particular languages.

References:

Apresjan, Ju.D. 2004. Akcional'nost' i stativnost' kak sokrovennye smysly (oxota na *okazyvat'*). In: Apresjan (ed.), *Sokrovennye smysly*. *Sbornik statej v čest' N.D. Arutjunovoj*. Moskva: Jazyki slavjanskix kul'tur. 13-33.

Apresjan, Ju.D. 2008a. O semantičeskoj motivirovannosti leksičeskix funkcij-kollokatov. *Voprosy jazykoznanija*, No. 5, 3-33.

Apresjan, Ju.D. 2008b. Anglijskij tolkovo-kombinatornyj slovar'. I. Leksičeskie funkcii. In: A.V. Bondarko, G.I. Kustova, R.I. Rozina (eds.), *Dinamičeskie modeli. Slovo. Predloženie*. *Tekst*. *Sbornik statej v čest' E.V. Padučevoj*. Moskva: Jazyki slavjanskix kul'tur. 20-58.

Apresjan, Ju.D. & M.Ja. Glovinskaja 2007. Two Projects: English ECD and Russian Production Dictionary. In: Kim Gerdes, Tilmann Reuther, Leo Wanner (eds.), *Meaning—Text Theory 2007*. *Proceedings of the 3rd International Conference on Meaning-Text Theory.* Wiener Slawistischer Almanach, Sonderband 69. München—Wien (to appear).

Iordanskaja, L.N. 1970. Popytka leksikografičeskogo tolkovanija gruppy russkix slov so značeniem čuvstva. *Mašinnyj perevod i prikladnaja lingvistika*, vyp. 13, 3-26.

Iordanskaja, Lidija & Igor Mel'čuk. 2009. Well-formedness of Semantic Structures (in this volume).

Mel'čuk, I.A. 1974. *Opyt teorii linguističeskix modelej "Smysl ⇔ Tekst".* Moskva: Nauka.

Mel'čuk, I.A. 1995. *Russkij jazyk v modeli "Smysl ⇔ Tekst".* Moskva—Vena: Wiener Slawistischer Almanach, Sonderband 39.

Mel'čuk, Igor A. & Alexander K. Zholkovsky. 1984. *Explanatory Combinatorial Dictionary of Modern Russian. Semantico-Syntactic Studies of Russian Vocabulary*. Vienna: Wiener Slawistischer Almanach, Sonderband 14.

Mel'čuk, Igor & Leo Wanner. 1996. Lexical Functions and Lexical Inheritance for Emotion Lexemes in German. In: L. Wanner (ed.), *Lexical Functions in Lexicography and Natural Language Processing*. Amsterdam-Philadelphia: John Benjamins Publishing Company. 209-278.

Mel'čuk, Igor et al. 1992. Mel'čuk Igor avec Nadia Arbatchewsky-Jumarie, Lidija Iordanskaja et Suzanne Mantha. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques III*. Montréal: Les Presses de l'Université de Montréal.

Mel'čuk, Igor et al. 1999. Mel'čuk Igor avec Nadia Arbatchewsky-Jumarie, Lidija Iordanskaja, Suzanne Mantha et Alain Polguère. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques IV*. Montréal: Les Presses de l'Université de Montréal.

Paducheva, E.V. 1991. Predikatnye imena v leksikografičeskom aspekte. *Naučno-tekhničeskaja informacija*, Serija 2, № 5. 21-31.

Reuther, Tilmann. 1994. O perifrastičeskix naimenovanijax rečevoj dejatel'nosti. In: N.D. Arutjunova (ed.), *Logičeskij analiz jazyka: Jazyk rečevyx dejstvij*. Moskva: Nauka. 76-82.

Reuther, Tilmann. 1996. On Dictionary Entries for Support Verbs: the Case of Russian VESTI, PROVODIT' and PROIZVODIT'. In: L. Wanner (ed.), *Lexical Functions in Lexicography and Natural Language Processing*. Amsterdam-Philadelphia: John Benjamins Publishing Company. 181-208.

Reuther, Tilmann. 2003. Support verb combinations with existential verbs (German and Russian). In: Sylvaine Kahane and Alexis Nasr (eds.), *Proceedings*. *First International Conference on Meaning-Text Theory / Première conférence internationale sur la Théorie Sens-Texte.* Paris: Ecole Normale Supérieure. 1-10.

Uspensky, Vladimir. 1979. O veščnyx konnotacijax abstraktnyx suščestvitel'nyx. *Semiotika i informatika*, odinnadcatyj vypusk. 142-148.

Concession in Russian: Semantics as a reflection of Rhetoric

Valentina Apresjan
Russian Language Institute
Volkhonka 18/2
Moscow Russia
valentina.apresjan@gmail.com

Abstract

The paper examines how the rhetorical functions of the "system-forming meaning" of concession determine its semantic properties. The paper suggests and motivates the definition of the core meaning of concession as reflected in the Russian conjunction *hotia* 'although'. Further analysis reveals the major semantic modifications to this core meaning, namely, the 'hypothetical', 'want' and 'quantity/degree' modifications. The paper shows how they combine and interrelate in the meaning of Russian concessive words and constructions.

1 Credits

This paper was written with the financial support of the following grants: RF President grant for the leading scientific schools NSH-3205.2008.6, RHSF 07-04-00-202a for "System-forming meanings of Russian Language", and the grant of the Program for Fundamental research of the Department of Humanities of the Russian Academy of Sciences «Russian culture in world history».

2 Introduction

Contemporary semantic theory offers many instruments of analyzing semantics in a systematic way. One of them is the notion of a system-forming meaning, introduced by Ju.D.Apresjan in (Apresjan, 2001; Apresjan, 2006). System-forming meaning is defined as "a meaning which constitutes part of many linguistic items of different nature (manifold form)… yet has consistent manifestations with regard to the rules of semantic interaction (projection rules)" (Apresjan 2006:52). Some examples of system-forming meanings include 'to know' and 'to think', which distinguish between *factuals* and *putatives* – two semantic categories comprised of linguistically diverse items which are nevertheless characterized by certain uniform properties. Some other system-forming meanings include 'goal', 'cause', 'condition' 'quantity', 'degree'. In some ways the notion of system-forming meaning reminds of Wierzbicka's semantic primitives – indeed, 'to know' and 'to think', as well as many other system-forming meanings suggested in (Apresjan, 2006:52), such as 'to do', 'can', 'to be', 'one' are included in the list of semantic primes (Wierzbicka, 1972). However, some other system-forming meanings are semantically more complex, e.g., 'goal', 'moment', 'to begin', 'to stop', and can be further decomposed into several components of the semantic primitives level.

'Concession' is also a system-forming meaning, expressed by numerous linguistic items, lexemes, constructions, phrasemes (*although, even though, despite, in spite of, yet, however, still, nevertheless, to concede, compromise, Try though he might, against all odds*, etc.), though one even more complex than 'cause', 'condition' or even 'goal' and therefore not as linguistically and typologically all-pervading or universal. Yet it is also one of the very important and central notions, both linguistically and cognitively. The cognitive operation "served" by the linguistic category of concession, is the perception of two facts as discordant, contradicting each other and, therefore, violating certain natural norms or expectations. When we say *Although it was warm outside, she put on a fir coat*, we logically link two separate facts and

thus perceive the existing state of affairs as abnormal, unexpected or unusual (the normal, expected situation would be to dress lightly in warm weather). In this sense concession is as important a category as 'cause' or 'condition' – though it is semantically more complex, as, unlike the latter two, it establishes not a natural link between two facts, but rather a "glitch" in those links. Semantically concession is usually defined in terms of condition (or, in some accounts, causality), combined with negation – 'usually if X, Y; in this case, X and not-Y' or 'usually X results in Y; in this case, X and not-Y' (Grevisse, 1969; Liapon, 1986; König, 1986; König, 1991; Morel, 1996; Khrakovskij, 1998]. However, this accounts only for the main body of concessives, such as *although, despite, in spite of*. As a systematic study shows, concession is related to certain other system-forming meanings, besides condition and negation, both on the contemporary and historical levels. Moreover, as will be shown below, these other semantic ties are motivated, given the rhetoric functions of concessives in an utterance. The current paper strives to analyze the lexical and grammatical category of concession systematically, in terms of system-forming meanings, and to show how several simpler system-forming meanings interact and combine to give rise to this more complex concept. Another goal of this paper is to bring together two approaches to analyzing concessives – one that may be called logical and another that may be called rhetorical.

The term "concession" itself (from *concessio*), which was introduced in Latin grammars reflects the *rhetorical* approach in which concession was viewed not as a meaning or a group of meanings verbalized in certain words and expressions, but rather as a special figure of speech. Thus, Diderot's "Encyclopedie" gives the following definition to concession (Diderot: 804): "CONCESSION, s. f. figure de Rhétorique par laquelle l'orateur, sûr de la bonté de sa cause, semble accorder quelque chose à son adversaire, mais pour en tirer soi-même avantage, ou pour prévenir les incidens inutiles par lesquels on pourroit l'arrêter".[1] As we can see, concession as a figure of speech in this account is not necessarily associated with the use of special lexical or grammatical means – linguistic concessives. On the other hand, in the contemporary linguistic thought concessives are treated as a distinct semantic and grammatical category, expressed by various syntactic and lexical means. Reflecting this new tendency, Grammaire Française 1967 defines the semantics of concession as a certain conflict which consists in "bringing together two facts which normally exclude one another", as quoted after Morel (1996).

However, these two approaches do not contradict each other; in fact, quite the opposite is true: one can understand the semantics of concessives better if one looks at their rhetoric function in an utterance. Rhetorically, concession is a figure of speech, an instrument of advancing one's own point of view in a subtle fashion. The speaker accomplishes that by seemingly acknowledging (to an extent) the right of the opponent in order to emphasize his/her own viewpoint more strongly. The more one concedes, the more one needs to win back. Rhetorically, using concessives is like saying to an opponent: "I admit that part or all of what you are saying might be true; but even in the case it is true, even if all of it is true, it does not change the situation; I still want to make a statement to the contrary, or a statement that would weaken your viewpoint, and I believe I am right".

What are the main components of this rhetoric frame, if we analyze it semantically?

1) first of all, it is the condition – 'if'; 2) next, it is the admittance of the opponent's right. Admittance is a kind of agreement given against one's wishes. The speaker admits something unwanted, yet wishes to advance his/her own point of view. Hence, the next important component – 'want'; 3) another semantic component is 'true' – since the speaker acknowledges, to an extent, the truth of his/her opponent's opinion; 4) there is also a component of 'probability', facts versus suppositions, as the speaker is willing to admit more than might actually be taking place at the moment – hence, the factual and putative components 'is' and 'may'; 5) finally, there is a component of an overstatement; the speaker wants to make his/her point very strong, and (s)he adds this credibility and polemic strength by stating that even though *all* of what the opponent thinks might be true, this still does not change the situation. This triggers another semantic component – that of extreme 'degree' or 'quantity'. The above semantic components are relevant for concessive words and constructions, both on the synchronic and diachronic levels.

Diachronically, many concessives evolved from linguistic items originally expressing admittance,

[1] "Concession: a figure of speech whereby the orator, sure of his own right, outwardly partly agrees with his opponent in order to gain advantage or to thwart unnecessary objections by which he could be stopped".

agreement, truth, wish, degree, quantity. Consider, for example, the following Russian data: 1) concessive conjunctions *hotia* and *hot'* which are etymologically related to the verb *hotet'* 'want'; 2) concessive conjunction *pravda* which has rather recently evolved from the noun *pravda* 'truth'; 3) agreement particle *konechno* 'of course' in its concessive usage (*On, **konechno**, paren' neglupyj, no ochen' lenivyj* '**Granted**, he isn't stupid, but he's excessively lazy'); 4) concessive conjunction *tol'ko* which evolved from the quantifier *tol'ko* 'only'; 5) concessive phrasemes *tol'ko by* and *lish' by* 'if only', which also come from the quantifier 'only'; 6) concessive phraseme *malo li* 'lit.: little whether' + *wh*-word (***Malo li** chto emu v golovu vzbredet, nel'zia co vsem soglashat'sia* '**Whatever** crazy idea might go into his head, you cannot agree with everything'); 7) probability modal *mozhet* 'may' in its concessive usage as a factual parenthetic word (*Ja, **mozhet**, tozhe kandidat nauk, a ne zadajus'* 'I **am** also a Ph.D., but I don't give myself airs').

Synchronically, the semantic components of 'condition', 'want', 'extreme quantity/degree', 'correspondence to reality', 'probability', 'agreement', in different combinations, form the semantic core of concessive words, as well as various semantic additions to it, which constitute the more marginal parts of the semantic field of concession. Below is the semantic analysis of Russian concessives, lexemes, phrasemes and constructions. Because of space limitations, we are unable to consider the entire corpus of concessive data, yet what is presented below gives a fairly representative account of the how this field is semantically organized.

3 The semantic core of concession

Below is the semantic analysis of Russian concessives, lexemes, phrasemes and constructions. In Russian, the semantic core of concession is expressed by the conjunction *hotia* 'although'. It embodies the concessive meaning in its purest, without any additional semantic components. Even its closest synonym *xot'* 'although' has an additional stylistic component – it belongs to the colloquial register. Another close synonym of *hotia*, the preposition *nesmotria na* 'in spite of', is somewhat stronger that *hotia* and thus more limited in the range of available contexts. Therefore, we will consider *hotia* and then show how its meaning becomes enriched and complicated by additional components in the meanings of other Russian concessives. There are two semantic approaches to analyzing concessive meaning – either through the combination of causality and negation (Liapon, 1986; König, 1991; Uryson, 2003), or through the combination of condition and negation (Grevisse, 1969; König, 1986; Morel, 1996; Khrakovskij, 1998][2]. Consider, e.g., the following definitions of the English conjunction *although* in König (1986) and (1991):

(1) *although p, not q*
'a. p and not-q;
b. if p, usually q'

(2) *not (q because p) = (not-q) although p*
This house is no less comfortable because it dispenses with air-conditioning = This house is no less comfortable although it dispenses with air-conditioning.

As we can see from definition in (2), in König's analysis *because* under wide-scope negation equals *although* under narrow-scope negation; it means that the situation P, introduced in the concessive part, is a "failed reason" for not-Q, one that turned out to be insufficient to thwart Q. Of these two definitions, the "if"-based one appears to be semantically more justified. In fact, the following definition for *hotia* 'although' is suggested (V. Apresjan, 2006:631).

[2] A definition of the Russian *hotia* 'although', which does not subscribe either to the conditional or to the causal account is given in (Iordanskaya, Mel'čuk, 2007:481), where *hotia* is analyzed as the antonym of *raz* 'since' and is given the following definition: *Hotia Q, P* 'pri nalichii fakta Q fakt P <proiznesenie vyskazyvanija P> javliaetsja neestestvennym' ('in the presence of fact Q, fact P <pronouncing the utterance P> is not natural'). Thus, the relation between the two facts is seen not in terms of a failed dependence, but in terms of co-existence.

(3) *Hotia P, Q* (*Hotia on byl bolen* [P]*, on poshel na rabotu* [Q] 'Although he was ill, he went to work') =

'P is taking place; Q is taking place; the speaker thinks that if something like P is taking place, usually something like not-Q is taking place'.

First of all, it is important to clarify the component 'speaker thinks'; it means that concession contains a reference to certain universal tendencies according to which the co-existence of P and Q is strange. However, these tendencies exist "in the eye of the beholder", they are not necessarily objective. In this respect, 'concession' is different from 'condition' or 'cause' which do not contain such implications. Compare the following pair of examples, correct (4) VS. the pragmatically inappropriate (5):

(4) *It is strange to get so upset because of a card loss*

(5) ??*It is strange to get so upset despite a card loss*

The causal conjunction *because* does not contain a reference to the speaker's opinion; therefore, the phrase is grammatical. The concessive preposition *despite* contains a reference to the speaker's opinion that there is a general tendency according to which people get upset when they lose money at cards. Therefore, the phrase contains *two* conflicting speaker's opinions – one in the assertion, that it is *strange* to get upset after a money loss, and another in the presupposition, that it is *normal* to get upset after a money loss (introduced into the utterance by the preposition *despite*). Another point in need of clarification in the definition of *hotia* is the use of semantic components 'something *like* P', 'something *like* Q' rather than referring directly to the situations P and Q, which are taking place. However, often the interdependence exists not between P and not-Q, but between some other situations whose existence can be somehow deduced from the existence of P and Q; cf.:

(6) *Hotia my s nim kazhdoe voskresen'je p'jem vmeste vodku, v dolg ia u nego poprosit' ne mogu*
'Although we drink vodka together every Sunday, I can't ask to borrow money from him'

Obviously here, the tendency to which the speaker is implicitly referring is not 'If we drink vodka every Sunday, usually I can ask to borrow money' but rather a more general one 'If people are friends, usually they can expect help from one another'. In this sentence, the existence of friendship is deduced from drinking together, and the borrowing of money is viewed as a particular instance of help. As concerns the component 'if', rather than 'because' in the definition, let us consider the following examples to prove the former is preferable:

(7) *Although he achieved fairly good marks at Harvard* [P]*, he is definitely not an academic type* [Q]

(8) *Although the forecast was bad* [P]*, the day turned out to be sunny* [Q]

If we try to analyze these examples in terms of failed causality, we will get semantically implausible interpretations such as 'Achieving good marks could, but didn't lead to the fact that he is an academic type'; 'A bad forecast could, but didn't lead to the fact that the weather was bad'. The conditional account, on the contrary, produces natural interpretations 'If one achieves good grades, it is natural to expect that person to be an academic type'; 'If the forecast is bad, it is natural to expect bad weather'. Consider also the following Russian example, where causality analysis fails to provide a semantically and pragmatically plausible interpretation:

(9) *Hotia ego familija Ivanov, on ne russkij*
'Although his last name is Ivanov, he's not Russian'

Likewise, the fact that one's last name is Ivanov does not *lead* to the fact when one is Russian; the dependence here is conditional, not causal: "If one's last name is Ivanov, it is natural to expect such a person to be Russian'.

3.1 Modifications to the semantic core of 'concession'

The basic semantic core of concession can be modified to give rise to many more concessive items. These modifications take place in the form of adding semantic components related to the following system-forming meanings: 'may' (hypothetical character of situations); 'quantity/degree'; 'want'. Another mechanism of modifying the core concessive meaning is conversion, which gives rise to a whole large class of concessives (traditionally considered contrastives) – such as *vse-taki, vse zhe, vse ravno, tem ne menee* 'still', 'nevertheless', 'and yet'. In fact, these words describe the same situation as the traditional concessives like *although*, only from a different perspective. Cf. (10) VS. (11)

(10) *My poshli guliat', hotia shel dozhd'*
'We went for a walk [Q] although it was raining [P]'

(11) *Shel dozhd', no my vse-taki poshli guliat'*
'It was raining [P], yet we went for a walk [Q]'

These words differ from the prototype of concession in the following: in the semantic formula 'P is taking place; Q is taking place; the speaker thinks that if something like P is taking place, usually something like not-Q is taking place' they, unlike the traditional concessives *hotia* 'although' and *nesmotria na* 'despite', introduce the situation Q. However, semantically, they are not different from the traditional concessives[3]. Conversion is semantically the most trivial of modifications; below are considered semantic additions to the core concessive meaning, along the above-mentioned semantic lines - 'may' (hypothetical character); 'quantity/degree'; 'want'.

3.2 Modification of 'hypothetical character'

The semantic core of concession is factual, that is, situations P and Q are actually taking place, as reflected in the definition 'P is taking place; Q is taking place; the speaker thinks that if something like P is taking place, usually something like not-Q is taking place'. However, the components 'P is taking place' and 'Q is taking place' can be modified. Let us consider the concessive conjunction *pust'* 'let' in phrases like

(12) *Pust' veter, pust' buria, ja nichego ne ispugajus'*
'Let there be wind, let there be storm, I will not be afraid of anything'

(13) *Pust' prosit, pust' umoliaet, ja ego vse ravno ne proshchu*
'Let him ask, let him implore, I will not forgive him'

As can be seen from these examples, *pust'* (same as its close colloquial synonym *puskaj* 'let') in contrast to the factual *hotia*, refers to prospective and hypothetical situations. The following definition is proposed for *pust'* and *puskaj*:

(14) *pust' <puskaj> P, Q* (*Pust' ugrozhaet* [P]*, ja ne pojdu na etu aferu* [Q] 'Let him threaten me, I will not get involved in this racket'; *Pust' on luchshij uchenik v klasse* [P], *u nego net tvorcheskogo*

[3] Here, and elsewhere we speak about the primary meaning of the words *hotia, vse-taki* etc.; most of them have other meanings, sometimes quite numerous (as the conjunction *hot'*); yet unless specified otherwise, it is the primary concessive meaning that we consider.

myshlenija [Q] 'He might be the best student in class, he lacks creativity') = 'The speaker admits that P might be taking place now or might take place in the future; the speaker thinks that Q is taking place or will take place in the future; the speaker thinks that if something like P is taking place, usually something like not-Q is taking place'

Interestingly, the modification along the lines of hypothetical situations frequently "invites" the semantic addition of degree; the hypothetical situations are often also characterized by extremeness – the situations tend to be of enormous proportions:

(15) ***Puskaj*** *ja umru pod zaborom kak pios, /* ***Pust'*** *zhizn' menia v zemliu vtoptala, / - Ja veriu: to Bog menia snegom zanios, / To buria menia celovala!* (A.Blok, Poets)
'**Let** me die in the gutter like a dog/ **Let** life tramp me into the ground/ I believe – it was God Who covered me with snow/ It was the storm that has kissed me'

This semantic property is corroborated by the co-occurrence properties of these conjunctions: *pust'* and *puskaj* tend to co-occur with verbs denoting "extreme" or destructive actions and happenings, such as *to kill, to die* etc. (*Pust' ja umru* 'Even if I die'; *Puskaj menia ub'jut* 'Let them kill me'), as well as with nouns denoting big quantity and adjectives denoting high degree (*Pust' xot' million mne predlozhat* 'Even if they offer me a million', *Puskaj on samyj luchshij specialist* 'Even if he is the best specialist'). This is true not only of *pust'* and *puskaj*, but also of other concessives that contain the "hypothetical" modification, namely, *hot', hot' by, hotia by, dazhe esli, pust' dazhe* 'even if'*, tol'ko by, lish' by* 'if only'*, esli uzh ne…to* 'if not, then'*, po krajnej mere, hot', hotia by* 'at least'. In some of these concessive items hypothetical situations are of enormous proportions, in some they are just the opposite – practically negligible, yet there is always a kind of extremeness, a kind of polarity present. This is a not an unexpected extension of meaning: after all, if a situation is hypothetical, it is natural to explore it to the extreme and consider if things might be different should this situation be taking place in a very large or a very small degree. Or perhaps, the extension of meaning is induced the other way round: hyperbole and litotes belong to the hypothetical world rather than the world of reality, therefore, when they appear in reference to some situations, these situations automatically revert to the hypothetical status.

Below is the analysis of other "hypothetical" concessives. Some of the above-listed concessives will be considered in the next two sections, under the 'quantity/degree' or 'want' headings, depending on which modification is more crucial to their meaning: e.g. the items *tol'ko by, hot' by, lish' by* which, although they also contain the 'hypothetical' and 'degree' components, are mostly centered around the 'want' constituent. On the other hand, *hotia by, hot'* [4][5]*, po krajnej mere*, although they also point both to the 'hypothetical' and 'want' modifications, are mostly about 'quantity/degree'. Consider the following examples with concessive items meaning 'even if' (synonymous to *pust'* and *puskaj*), where hypothetical character of the situations is combined with a hyperbole:

(16) *Nichego ne skazhu, hot' rezh'te menia*
'I will not say anything even if you cut me to pieces'

(17) *Pust' dazhe v tiur'mu menia posadiat, ja ne postuplius' svoimi principami*
'Even if they put me to jail, I won't give up my principles'

(18) *Dazhe esli luchshego mastera prishliut, on s etim remontom ne spravitsia*
'Even if they send the best craftsman, he won't be able to manage these repairs'

[4] Most of the items considered are polysemous, that is why they appear under more than one heading; *hot',* for example, has eight meanings, four of which are concessive.

[5] This concerns *hot'* in the meaning of 'at least'; there is another *hot*', in phrase (16), which means 'even if' and is mostly centered around the 'hypothetical' modification.

(19) *Esli i golodat' budu, vse ravno svoboda vazhnee*
'Even if I have to starve, freedom is still more important'

(20) *Hotia by i kvartiru prishlos' prodat', a s dolgami rasplatit'sia nado*
'Even if it's necessary to sell the apartment, debts must be paid'

(21) *Da hot' by tam potop byl, mne nado na vstrechu*
'Even if there were a flood there, I'll have to go to the meeting'

Pust' dazhe, dazhe esli, esli i, hotia by, hot' by have similar meanings, which are explicated as follows:

(22) *Pust' dazhe <dazhe esli, esli i, by, hot' by> P, Q* = 'The speaker admits that a very considerable P might be taking place now or might take place in the future; the speaker thinks that Q is taking place or will take place in the future; the speaker thinks that if something like P is taking place, usually something like not-Q is taking place'

Like *pust'* and *puskaj*, these expressions tend to co-occur with the denotations of extreme and destructive actions and events, with big quantity nouns and high degree adjectives; cf. the phrases above. The concessive conjunction *hot'* (not in its primary meaning of 'although', but in the meaning of 'even if') is semantically close to them, it also contains the 'hypothetical' and 'quantity/degree' modifications, yet it usually describes the impossibility to achieve the desired situation despite one's considerable efforts:

(23) *Hot' darom otdavaj, nikto u tebia etu mashinu ne voz'met*
'Even if you give it out for free, nobody will take this car'

(24) *Hot' na koleniax umoliaj, on tebe ne pomozhet*
'Even if you implore him on your knees, he won't help you'

Hence, the following definition is proposed for *hot'* in this meaning:

(25) *Hot' P, Q* = 'The speaker is sure that even if a very considerable P takes place, the desired Q will not take place; the speaker thinks that if something like P is taking place, usually something like not-Q is taking place'

Usually combined with a verb in the imperative, *hot'*, like the above-described concessives, also co-occurs with verbs meaning destruction, as well as expressions denoting great quantity or high degree, often phraseological: *hot' umri* 'even if die', *hot' veshajsia* 'even if hang oneself', *hot' semi piadej vo lbu* 'even if wise as Solomon'. *Hot'* itself forms a number of phrasemes with a rhetorical hyperbole, meaning 'whatever one does, the goal will not be achieved' (English phrasemes 'not for toffee', 'can't for the life of me'): *hot' rezh'* 'even cut', *hot' ubej* 'even kill', *hot' zarezh'* 'even stab', *hot' tresni* 'even hit'. Cf. also *hot' kol na golove teshi* 'even pole on head char' with the meaning 'X can't understand something for the life of X'.

3.3 Modification of 'want'

One more modification to the semantic core of concession is that of 'want' and, like that of 'degree', it is often combined with the 'hypothetical' modification. Again, this is something to be expected: after all, a *desired* situation is usually one which has not yet been attained; thus, it is *hypothetical* rather than *factual*. Likewise, 'want' is frequently accompanied by 'quantity/degree' modification: one can *strongly* wish something to happen, one can be prepared to give up *anything* for the desired object, one can be satisfied with a *small* part of what is desired, etc. Considering the core concessive meaning 'P is taking place; Q is taking place; the speaker thinks that if something like P is taking place, usually something like

not-Q is taking place', there are two components which the modification of 'want' can affect. Namely, the situation Q can be viewed as desirable, and the fact that it might take place can be conjectured as a certain victory over the unfavorable circumstances, often at the expense of letting the undesirable situation P to take place. There are several classes of concessives which incorporate the 'want' modification. One class, which is perhaps the most representative, can be grouped under the heading 'if only'. It includes such items as *hot' by, tol'ko by, lish' by* 'if only'; cf. *Hot' by uspet'* 'Let me be on time'; *Tol'ko by on prishel* 'Let him come'; *Lish' by ona byla zdorova, ostal'noe nevazhno* 'If only she were healthy, the rest doesn't matter'. Of these three, *hot' by* is a "truncated" concessive because it is mostly used to express pure wish rather than concession, although concessive usage is also possible; cf. the relatively much higher frequency of optative phrases like *Hot' by otdohnut' nakonec!* 'If only I could finally have some rest' in comparison to concessive phrases like *Hot' by ne meshal* [P]*, esli uzh pomoch' ne mozhesh'* [Q] 'You could at least stop pestering me if you cannot be of help'[6]. The following definition is proposed for *hot' by*:

(26) *Hot' by P* 'The speaker wants P; perhaps the speaker wants a greater Q but thinking that Q cannot be had, (s)he says (s)he wants to have P'.

Tol'ko by and *lish' by* can also be used as optatives, yet they are somewhat more typical in concessive usages[7]:

(27) *Ej samoj nichego ne nuzhno, tol'ko by deti byli v poriadke*
'She doesn't require anything so long as her children are OK'

(28) *On gotov na liubuju podlost', lish' by otomstit'*
'He's ready to commit any kind of foul trick in order to get revenge'

The following definition is proposed for *tol'ko by* and *lish' by*:

(29) *Pust' Q, tol'ko by <lish' by> P* 'Let there be Q, if only P' = 'The speaker wants P very much; thinking that the desired P might be accompanied by an undesirable Q, the speaker says (s)he is ready for Q to take place and wants P to take place'

Another class of concessives affected by the 'want' modification is represented by particles which can be grouped under the semantic heading 'at least': *hot', hotia by, po krajnej mere*. These represent a scale, both in terms of the wish's intensity and the quantificational characteristics of the desired object/situation. The lower the demands, the stronger the wish. Thus, the particle *hot'*, which implies very low hopes on the part of the experiencer and the readiness to be satisfied with the least available resources, also points to the strongest desire: *Podari mne hot' mgnovenie schast'ja!* 'Give me at least one moment of happiness'; *Hot' na minutu ostav' menia v pokoe* 'Leave me alone at least for one minute'. It usually co-occurs with denotations of small quantity or even negative polarity items: *hot' mgnovenie* 'at least an instant', *hot' glotok vody* 'at least one gulp of water', *hot' na jotu sochuvstivija* 'at least a jot of sympathy', *hot' na gramm ponimanija* 'at least an ounce of understanding', etc. *Hotia by* occupies the intermediate position, and can be used to voice minimalistic, as well as reasonable requests. *Po krajnej mere* is the least emotionally loaded item and implies a certain sufficient minimum of resources rather than their smallest possible amount. It can be used not only in requests, but also in demands: *Vy dolzhny po krajnej mere obespechivat' bezopasnost' uchashchihsia* 'You must at least guarantee the students' safety'; cf. the impossibility of this phrase with the particle *hot'*: **Vy dolzhny hot' obespechivat'*

[6] According to the data from the Russian National Corpus, approximately one phrase out of fifty in which *hot'* is used in this meaning, is concessive, with the rest being optative.

[7] According to the data from the Russian National Corpus, approximately one third of the examples reflects optative usage, with two thirds being concessive.

bezopasnost' uchashchihsia. The following definitions are suggested for these items (the varying part is in bold):

(29) *Hot' P* 'Thinking that the desired Q is impossible, the speaker or the subject **wants** a **very small** P from the same class of situations as Q'

(30) *Hotia byt P* 'Thinking that the desired Q is impossible, the speaker or the subject **is ready to be satisfied** with a **small** P from the same class of situations as Q'

(31) *Po krajnej mere P* 'Thinking that the desired Q is impossible, the speaker or the subject **accepts** a **smaller** P from the same class of situations as Q'

The smallest degree and the strongest desire also trigger the least probable situation: while *hotia by* and *po krajnej mere* can be used in past contexts, where P is a fact, for *hot'* such usage is restricted; it is mostly used with imperatives, in contexts of request: *Hot' kusok hleba podaj!* 'Give me at least a piece of bread!'.

3.4 Modification of quantity/degree

As noted above, the modification of 'quantity/degree' often accompanies 'hypothetical' and 'want' modifications. It is difficult to say which is primary in each case. Certainly, it is not difficult to give meaning to the fact that situations characterized by extremely large or small quantity/extremely high or low degree are often hypothetical; after all, extreme situations are not as frequent in everyday life as average ones, therefore when they are mentioned, it is usually in the hypothetical, rather than factual, context. However, there are items where the 'quantity/degree' modification is not necessarily accompanied by 'hypothetical' or 'want' modifications. One of such items is the syntactic construction consisting of a *k*-pronoun[8] (*wh*-word) + negative particle. This pragmatically "pessimistic" construction implies that despite numerous circumstances P conducive to the situation Q, it did not take or will not take place; cf.

(32) *Kak on ni staralsia, a nichego u nego ne vyshlo*
'Try though he might, ne didn't succeed'

(33) *K kakim advokatam on ni obratitsia, delo on vse ravno proigraet*
'Whatever attorneys he'll solicit, he'll nevertheless lose the case'

A syntactic construction very close in meaning to *wh*-word+*ni* is *pri vsem P* 'with all P' ('all P notwithstanding'). It also implies that despite all the circumstances in favor of Q, it did not or will not take place; cf.

(34) *Pri vseh svoih nesomnennyh dostoinstvah, on ne podhodit na etu dolzhnost'*
'All his indisputable qualities notwithstanding, he's not suited for this post'.

The following definition is proposed for *k*-pronoun + *ni*... and *pri vsem P* constructions:

(35) *Kak ni P, not-Q*; *pri vsem P, not-Q* 'a very considerable P is taking place; Q is not taking place; the speaker thinks that usually when something like P is taking place, something like Q is taking place'.

Compared to the core concessive meaning 'P is taking place; Q is taking place; the speaker thinks that if

[8]Such pronouns as *chto* 'what', *skol'ko* 'how much', etc., though they do not formally start with *–k* are considered *k*-pronouns, in the same vein as the notion of *wh*-word is extended to such items as *how*.

something like P is taking place, usually something like not-Q is taking place', this definition has the reverse distribution of negation. It reflects the fact that these constructions are inherently negative: despite many efforts aimed at Q, Q is not taking place. It is reflected in their co-occurrence properties, namely that sentences containing these constructions in the subordinate clause usually contain negation in the main clause. There is one more phraseme, namely *nesmotria ni na chto* 'against all odds', 'through thick and thin', 'no matter what', 'in spite of everything'. While the constructions *kak ni P* and *pri vsem P* are pragmatically "pessimistic", *nesmotria ni na chto* is "optimistic". Namely, it means that despite considerable obstacles, situation Q did take or is taking place. Though that is not obligatory, this phraseme often also refers to a great *wish* to achieve the desired situation. Thus, some of the typical contexts for this phraseme are *dobit'sia svoego nesmotria ni na chto* 'to achieve what one wants against all the odds', *verit' nesmotria ni na chto* 'to believe despite everything', *pobedit' nesmotria ni na chto* 'to win against all odds'. The valency P, that of an obstacle, is permanently filled in this phraseme with *ni na chto*, so only the valency Q, that of a situation, is free. The following definition is proposed for *nesmotria ni na chto*:

(36) *Q, nesmotria ni na chto* 'a very considerable P took place; Q is taking place; the speaker thinks that usually when something like P is taking place, something like not-Q is taking place'.

References

Apresjan, Jury. 2001. System-forming meanings 'to know' and 'to think' in Russian, *Russkii iazyk v nauchnom osveshchenii*. 2001. № 1. pp.5-26.

Apresjan, Jury. 2006. The foundations of systemic lexicography. *Linguistic picture of the world and systemic lexicography.* Iazyki slavianskix kul'tur. Moscow.

Apresjan, Valentina. 2006. Concession in language. *Linguistic picture of the world and systemic lexicography.* Iazyki slavianskix kul'tur. Moscow.

Grevisse, Maurice. 1969. *Le bon usage: Grammaire Française avec des remarques sur la langue Française d'aujourd'hui*. Neuvième édition revue. Éditions J. Duculot, S.A., Gembloux (Belgique), Librairie A. Hatier 8, Rue D'Assas, Paris 6e 1969.

Iordanskaja, Lidiya & Melčuk, Igor. 2007. *Smysl i sočetaemost' v slovare*. Iazyki slavianskix kul'tur. Moscow.

Khrakovskij, Viktor. 1998. Theoretical analysis of concessive constructions (semantics, assessment, typology). *The typology of concessive constructions.* Ed. by V. Khrakovskij. Nauka. Saint-Petersburg.

König, Ekkehard. 1986. Conditionals, Concessive Conditionals and Concessives: areas of contrast, overlap and neutralization. *On Conditionals*. Ed. By E. G. Traugott, A. ter Meulen, J. S. Reilly, C. A. Ferguson. Cambridge University Press. Cambridge, London, New York, New Rochelle, Melbourne, Sydney.

König, Ekkehard. 1991. Concessive Relations as the Dual of Causal Relations. *Semantic Universals & Universal Semantics*. D. Zafferer (ed.) Foris Publications. Berlin, New-York 1991. pp. 190-209.

Liapon, Maya. 1986. Semantic structure of a complex sentence and text. On typology of intertextual relations. Nauka. Moscow.

Morel, Mary-Annick. 1996. *La Concession en Français*. Editions Ophrys, 1996.

Uryson, Elena. 2003. Semantic and valency structure of words with concessive meaning. *Russkii iazyk v nauchnom osveshchenii*. 2003. № 6. pp.214-246.

Wierzbicka, Anna. 1972. *Semantic Primitives*. Frankfurt a. M.: Athenäum-Verl.

Constitution d'un corpus annoté autour du lexique des émotions: collocations et fonctions lexicales

Magdalena Augustyn
Laboratoire de Linguistique et Didactique des Langues Etrangères et Maternelles (LIDILEM), Université Grenoble 3
BP 25, 38040 Grenoble Cedex 09
Magdalena.Augustyn@u-grenoble3.fr

Agnès Tutin
Laboratoire de Linguistique et Didactique des Langues Etrangères et Maternelles (LIDILEM), Université Grenoble 3
BP 25, 38040 Grenoble Cedex 09
Agnes.Tutin@u-grenoble3.fr

Abstract

In this paper, we report an experiment of annotation of collocations in texts for pedagogical purposes using the Lexical Function model. We first show why showing collocations in context is according to us a fruitful method, and we present our annotation scheme and the corpora used. We then present some problems raised by the annotation process: delimitation of collocations, consistency of Lexical Functions, treatment of metaphors. All in all, the LF model appears to be operational, since over than 90% of collocations could be annotated with standard LFs in our corpus. We think that the model would probably benefit from being simplified (and homogenized) for a wider use for pedagogical purposes.

1 Introduction

Cette étude sur l'annotation des collocations dans le champ lexical des émotions s'inscrit dans le prolongement d'un projet plus global autour de l'étude des marqueurs linguistiques de la subjectivité, dans une perspective didactique[1]. Dans ce cadre, deux grands types d'études linguistiques ont été effectuées : d'une part, l'étude des marques du discours rapporté et des passages entre guillemets, autour des phénomènes de polyphonie (Rinck & Tutin, 2007) ; d'autre part, une étude autour du lexique des émotions, qu'il s'agisse du repérage et traitement sémantique des formes simples (Augustyn et al., 2008), sujet qui a également été traité à plusieurs reprises dans notre équipe (Goossens, 2005 ; Tutin et al., 2006), ou comme dans la présente communication, des collocations.

Notre objectif est ici d'utiliser et d'adapter le modèle des fonctions lexicales, afin d'annoter les collocations dans des corpus textuels. Les corpus annotés pourraient être utilisés à des fins didactiques, afin d'illustrer le phénomène collocatif dans son environnement « naturel », le texte. Nous avons souhaité exploiter les descriptions lexicographiques de la lexicologie explicative et combinatoire, en particulier la modélisation proposée par les Fonctions Lexicales, pour les projeter sur les occurrences textuelles, tout en souhaitant fournir des descriptions simplifiées pour des publics non spécialistes. Cependant, cette procédure s'est avérée plus complexe qu'envisagé et l'annotation textuelle a permis de mettre en évidence un ensemble de points problématiques que nous souhaitons exposer ici.

Après avoir présenté l'intérêt que représente pour nous l'élaboration de corpus annotés intégrant des informations phraséologiques, nous exposerons la méthodologie et les corpus utilisés. Nous aborderons ensuite les difficultés posées par l'annotation de ces phénomènes à l'aide du modèle des fonctions lexicales, et montrerons que le processus d'annotation permet d'enrichir la réflexion sur la modélisation des collocations.

[1] PPF piloté par le LIDILEM (2003-2007) (F. Grossmann et G. Antoniadis) : « Développement et exploitation de ressources linguistiques pour la didactique du français à l'aide d'outils de TAL. Etude des marqueurs linguistiques de la subjectivité et de la polyphonie. »

2 L'annotation des collocations dans les textes

Les associations lexicales décrites sous le terme de collocations posent de nombreuses difficultés, maintenant bien connues, aux apprenants en langue maternelle et étrangère (Cf. par exemple, Granger, 1998; Nesselhauf, 2005). Sans en examiner toutes les facettes, nous définirons la notion de collocation comme une association mémorisée de deux éléments linguistiques sémantiquement pleins qui entretiennent une relation sémantique directe, et généralement une relation syntaxique directe. Suivant l'analyse désormais classique de Hausmann (1978), nous distinguons dans ces associations binaires deux éléments au statut distinct : la base est l'élément stable de la collocation, et le collocatif est choisi contextuellement en relation avec la base. Dans le cadre de cette étude, nous avons choisi et adapté le modèle des fonctions lexicales de la lexicologie explicative et combinatoire (Mel'čuk et al., 1995), le modèle qui nous paraît le plus abouti dans la modélisation des collocations, afin d'annoter ces éléments lexicaux dans les textes.

Nous pensons en effet que, sur le plan didactique, l'utilisation de corpus annotés qui présentent la phraséologie dans un environnement textuel naturel est pertinente pour plusieurs raisons :

1. L'observation des collocations dans les textes permet une meilleure réutilisation des expressions. La plupart des didacticiens du lexique insistent sur la contextualisation du lexique pour l'apprentissage (par exemple, Tréville & Duquette, 1996 ; Cavalla & Labre, à paraître).
2. Les collocations étant généralement assez transparentes sur le plan sémantique, une mise en contexte pertinente est parfois plus éclairante pour les apprenants qu'une explication ou un métalangage complexes. Dans cette perspective, une réalisation remarquable est celle du dictionnaire des cooccurrences du logiciel Antidote (Charest et al., 2007), qui relie de façon systématique des collocations (appelées cooccurrences) à des exemples sélectionnés sur corpus. Un traitement sémantique systématique permet cependant aux apprenants de comparer les collocations et leur donne un accès onomasiologique, qui peut aussi être profitable sur le plan didactique.
3. L'observation sur corpus permet de mémoriser les propriétés syntaxiques des collocations (types de déterminants, tournures actives ou passives) qui doivent être apprises en même temps que les éléments du lexique, et que les dictionnaires mentionnent rarement (Tutin, 2004). Pour le champ qui nous concerne, on relève par exemple que des collocations comme *la peur paralyse* ou *le remords ronge* sont bien plus fréquentes au passif réduit (*paralysé par la peur, rongé de remords*) qu'à la voix active, information qui sera directement observable pour l'apprenant sur le corpus.

Malgré l'intérêt des corpus annotés comportant une information phraséologique, ce type de ressources n'a pas été énormément développé pour les collocations, à notre connaissance, hormis les expérimentations de Ludewig (2001), Fellbaum & Geyken (2005) et Tutin (2005)[2]. Cela est probablement dû à la difficulté de cette tâche qui exige un traitement extrêmement précis des objets lexicaux traités et ne peut en aucun cas être entièrement automatisé, comme nous le verrons.

3 Methodologie

3.1 Corpus

Le corpus utilisé pour notre étude comporte un ensemble de textes variés, principalement des écrits libres de droits puisque les corpus élaborés dans notre projet devaient être utilisables librement pour les usagers[3]. Cela restreint malheureusement le corpus à des œuvres plutôt anciennes. Notre objectif étant didactique, nous avons sélectionné pour les œuvres littéraires – qui constituent l'essentiel du corpus – des romans largement utilisés dans le cadre scolaire. Le corpus annoté, dont la composition détaillée apparaît dans le tableau 1, comporte ainsi des ouvrages classiques au programme des collèges comme *Les lettres de mon moulin* d'Alphonse Daudet, ou *La petite Fadette* de George Sand.

[2] Dans cette expérimentation, nous avions adapté un sous-ensemble des collocations du Dicouèbe de Polguère pour une annotation dans des textes littéraires. Ici, le corpus et le lexique traités sont nettement plus vastes.

[3] Rappelons que dans le droit français, les œuvres tombent dans le domaine public au bout de 70 ans.

Texte	Type	Nombre de mots
La petite Fadette (George Sand)	Littéraire	74456
Les Lettres de mon moulin (Alphonse Daudet)	Littéraire	46950
Colomba (Prosper Mérimée)	Littéraire	52619
Le petit chose (Alphonse Daudet)	Littéraire	83561
Le mystère de la chambre jaune (Gaston Leroux)	Littéraire	86 271
Contes (Perrault)	Littéraire	21684
Les contes du lundi (Alphonse Daudet)	Littéraire	68088
L'Ile mystérieuse (Jules verne)	Littéraire	199426
Le droit à la paresse (Jules Lafargue)	Essai polémique	12269
2 articles de la revue LIDIL	Ecrits scientifiques	10074
1 rapport scientifique *La place de la LSF dans l'intégration scolaire des enfants sourds* (Agnès Millet)	Ecrits scientifiques	13642
TOTAL		669040

Tableau 1. Composition du corpus annoté.

3.2 Ressources lexicales

Pour annoter les collocations au niveau du corpus, nous avons utilisé une procédure semi-automatique utilisant la base de données de collocations développée par Th. Fontenelle (1997)[4], que nous avons choisie du fait de sa grande couverture lexicale. Cette base de données, qui codifie les associations lexicales à l'aide du modèle des fonctions lexicales, a été constituée semi-automatiquement par Th. Fontenelle à partir du dictionnaire anglais-français Collins-Robert (pour la méthodologie utilisée, voir Fontenelle, 1997). Le tableau 2 ci-dessous donne un aperçu de la base utilisée.

Categorie syntaxique	Collocatif	Fonction Lexicale	Glose semantique
vt	Abîmer	Causdegrad	Dégrader
adj	Abject	Antiver	Mauvais
adj	Abominable	magn+antibon	intense et négatif
vt	Absorber	Causpredminus	faire diminuer
n	Absorption	s0causpredminus	Diminution
vt	Accabler	Nocer[5]	affecter vivement
vpron/vt	Accélérer	incepredplus/causpredplus	faire augmenter/augmenter
vt	Accentuer	Causpredplus	faire augmenter
n	Accès	Culm	Maximum
n	Accroissement	s0inceppredplus	Augmentation
vpron/vt	Accroître	causpredplus/inceppredplus	faire augmenter/augmenter
n	Accumulation	s0inceppredplus	Augmentation
vt	Accumuler	Causpredplus	faire augmenter
vt	Achever	culmreal1/real1	Réaliser

[4] Un très grand merci à Thierry Fontenelle de nous avoir autorisées à utiliser cette base lexicale, qui s'est révélée extrêmement riche. Nous regrettons que cette base n'ait pas davantage été exploitée dans le cadre des travaux sur la lexicologie explicative et combinatoire.

[5] Cette FL n'est pratiquement jamais utilisée dans le DEC ou le Dicouèbe.

Tableau 2. Liste de collocatifs associés aux noms d'affect extraite de la base de Fontenelle 1997 (et gloses correspondantes).

Nous avons extrait de cette base tous les collocatifs employés en cooccurrence avec les noms d'affect, avec l'indication de FL associée, ainsi qu'une glose de la fonction lexicale (cf. plus loin la discussion sur les gloses). Cette base, sous forme de table, a ensuite été appliquée à notre corpus en utilisant le système Intex développé par Max Silberztein (1998) et corrigé semi-automatiquement, en parcourant l'environnement lexical des noms d'affect. Les données du DEC (Mel'čuk et al., 1984, 1988, 1992, 1999) ainsi que celles du Dicouèbe ont également été consultées, mais leur utilisation n'a pas toujours été aisée pour les raisons que nous verrons plus bas.

3.3 Principes de base de l'annotation

Comme signalé plus haut, le modèle des Fonctions Lexicales nous paraît être un modèle riche pour le codage syntaxique et sémantique des collocations. Cependant, la modélisation apparaît évidemment complexe pour les non spécialistes que nous visons et nous avons souhaité la simplifier, tout en en conservant la philosophie, un peu à la façon du *Lexique Actif du Français* (Mel'čuk & Polguère, 2008), mais en poussant encore plus loin la simplification.

Les collocations sont annotées dans le corpus à l'aide du langage de balisage XML. La collocation est annotée sur le collocatif (élément <COLLOC>), la même base pouvant fréquemment être associée à plusieurs collocatifs, comme dans *avoir une peur terrible* (*avoir peur + une peur terrible*), comme on peut l'observer dans l'exemple (1).

Sur le collocatif, les éléments suivants sont annotés :

- Le **lemme de la base** (attribut BASE), qui permettra une recherche aisée dans le corpus, par exemple, pour obtenir tous les contextes où l'on a des collocations avec *amour*. Dans notre exemple, la base des collocations *avoir peur* et *une peur terrible* est *peur*.
- La **categorie syntaxique (et sous-categorie)** du collocatif (attribut CAT). On pourra ainsi rechercher toutes les collocations qui comportent un verbe transitif.
- La **fonction lexicale** (attribut FL) est également codée.
- Enfin, une **glose semantique** (attribut TYPE_SEM), devant permettre un accès onomasiologique à la collocation, est également proposée. Dans notre exemple, la glose pour *avoir* (*peur*) est /éprouver/ alors que la glose pour (*peur*) *terrible* est /intense et mauvais/.

(1) Pour le coup le petit Chose <COLLOC BASE="peur" CAT="vt" FL="Oper1" TYPESEM="éprouver">**eut**</COLLOC> **une** <LEXIQUE TYPE="affect" CAT="N" DOMAINE="peur" NV_LANGUE="courant" INTENSITE="moyen" POLARITE="négatif" >**peur**</LEXIQUE> <COLLOC BASE="peur" CAT="adj" FL="Magn+AntiBon" TYPESEM="intense et mauvais">**terrible**</COLLOC>; il se voyait déjà dans la rue, sans ressources... (*Le Petit Chose*, A. Daudet)

Sur la base (<LEXIQUE>), apparaissent de nombreuses informations sémantiques et syntaxiques concernant les mots d'affect : la catégorie, le champ sémantique, le niveau de langue, l'intensité, la polarité (pour une description détaillée, voir Augustyn et al., 2008). Précisons toutefois que le lexique de l'affect est traité dans notre corpus indépendamment des collocations.

Les principes de base de l'annotation étant posés, nous passerons maintenant aux difficultés rencontrées dans la mise en œuvre de l'annotation.

4 Problèmes rencontres dans l'annotation et solutions apportees

4.1 Les gloses des Fonctions Lexicales syntagmatiques

Le *Lexique Actif du Français* (Mel'čuk & Polguère, 2007) reprend les principes du *Dictionnaire Explicatif et Combinatoire* en les simplifiant et en les didactisant. Les associations lexicales, qu'elles soient syntagmatiques ou paradigmatiques, sont introduites à l'aide de « formules de description »

permettant un accès onomasiologique. Par exemple, dans l'article de EFFROI, les collocations *l'effroi prend, gagne, saisit* X est glosée de la façon suivante dans le LAF :

> **E. commence à être eprouve par X** envahir, gagner, prendre, saisir [N_X], s'emparer [de N_X]/[de l'âme de N_X]

Pour l'annotation, nous utilisons des gloses plus courtes, que nous supposons plus faciles à comprendre. Par exemple, pour la collocation précédente, nous utilisons la glose sémantique /envahir/, comme dans l'exemple suivant, tiré de *L'Ile Mystérieuse* :

(2) En effet, les singes, <COLLOC BASE="effroi" CAT="vt" FL="IncepFunc1" TYPESEM="envahir">**pris**</COLLOC> **d'un** <LEXIQUE TYPE="affect" CAT="N" DOMAINE="peur" NV_LANGUE="littéraire" INTENSITE="haut" POLARITE="négatif">**effroi**</LEXIQUE> subit, provoqué par quelque cause inconnue, cherchaient à s'enfuir. (*Ile mystérieuse*, J. Verne)

Les gloses sémantiques sont au nombre d'une cinquantaine, et apparaissent pour les collocations les plus productives. Elles sont moins précises sur le plan sémantique que les « formules de description » du LAF, mais la plupart des collocations étant assez transparentes sur le plan sémantique, nous pensons que le contexte permet de restituer facilement le sens. Dans notre démarche, nous privilégions ainsi la facilité d'accès au sens plutôt que la précision de la description, mais cette facilité d'usage doit être testée de façon concrète auprès d'usagers.

Pour obtenir un encodage plus homogène et respecter une certaine facilité de lecture, nous avons réduit les expressions à un système basé sur des étiquettes simples comme *commencer, intense, positif, négatif,* opérateur causatif *faire*, par exemple. Nous indiquons ci-dessous quelques-unes des correspondances rencontrées entre fonctions lexicales et gloses :

Glose semantique	**Fonction lexicale**
/commencer/	FL="IncepFunc0"
/commencer_éprouver/	FL="IncepOper1" ; FL="IncepPred"
/augmenter/	FL="IncepPredPlus"
/faire augmenter/	FL="CausPredPlus"
/affecter/	FL="Func1"
/affecter vivement/	FL="Magn+Fact1" ; FL="Magn+Func1"
/état/	FL="A1" ; FL="Adv1/2"
/intense/	FL="Magn"
/peu intense/	FL="AntiMagn"

Tableau 3. Exemple de correspondances glose/FL

Le problème des étiquettes sémantiques s'est aussi posé lors de l'annotation de fonctions non standard et de fonctions standard difficilement paraphrasables. Par exemple, nous avons recensé des collocations dans les textes où la fonction lexicale apparaît non standard, parce que la relation qui relie les éléments de la collocation n'est pas productive, par exemple : *liens d'amitié, crainte enfantine / puérile, amour maternel, amitié fraternelle*. Par ailleurs, même si la fonction lexicale est annotée, il nous est apparu difficile dans certains cas de trouver une glose simple, comme dans les cas suivants :

Oper1+Bon : *épanoui de Naffect*
NonOper1 : *soyez sans (inquiétude, crainte)*
A1Real2 : *stupéfait d'admiration*

Pour ces deux cas de figure, il aurait été possible d'attribuer les fonctions lexicales non-standard à la manière du Dicouèbe, mais dans notre système simplifié de codage, nous avons préféré signaler la

collocation, sans lui associer de traitement spécifique, comme dans l'exemple suivant pour *amitié fraternelle* :

(3) Peu à peu, quand il les vit honnêtes, énergiques, liés les uns aux autres par une <LEXIQUE TYPE="affect+relation" CAT="N" DOMAINE="affection" NV_LANGUE="courant" INTENSITE="moyen" POLARITE="positif" JUGEMENT="/" ATTITUDE="positif">**amitié**</LEXIQUE> <COLLOC BASE="amitié" CAT="adj" FL="" TYPESEM="">**fraternelle**</COLLOC> , il s'intéressa à leurs efforts. (*Ile mystérieuse*, J. Verne)

On peut cependant signaler que statistiquement, les fonctions lexicales non standard sont relativement peu nombreuses dans notre corpus (elles de représentent que 6,1% des collocations annotées). De la même façon, peu de collocations annotées ne reçoivent pas d'étiquette sémantique (seulement 4,4% de l'ensemble des fonctions standard), ce qui montre que notre système d'étiquettes sémantique est dans l'ensemble très couvrant.

4.2 Le problème de la delimitation des collocations dans les textes

Dans les textes bien entendu, les collocations ne se présentent pas sous la forme canonique qu'elles ont dans les dictionnaires. Les délimiter par des éléments XML n'est pas une tâche triviale et cette procédure doit suivre un ensemble de principes cohérents.

Le premier problème concerne les phénomènes de « mise en facteur » d'éléments de la collocation. Tout d'abord, une base peut être utilisée par plusieurs collocatifs, comme dans l'exemple (1), *avoir une peur terrible* (fusion de *avoir peur + peur terrible*). Ce cas de figure est traité assez simplement puisque la collocation est mentionnée sur le collocatif. Inversement – et moins fréquemment – il arrive qu'un collocatif porte sur plusieurs bases, comme dans l'exemple suivant, *il pleurait de rage et de désespoir*, où les collocations *pleurer de rage* et *pleurer de désespoir* sont fusionnées. Dans ce cas, la disjonction de la base sera indiquée dans un attribut de l'élément XML <COLLOC> (BASE="rage/désespoir").

Une autre question à traiter est la délimitation du collocatif même, lorsqu'il apparaît dans des formes composées, qu'il s'agisse de formes pronominales, de temps composés ou de formes passives. Nous avons décidé de n'intégrer dans l'élément du collocatif que la forme lexicale pleine, à l'exclusion des mots purement grammaticaux. Ainsi, pour les formes pronominales, les pronoms réfléchis ont été intégrés dans le collocatif pour les verbes intrinsèquement pronominaux comme *se pâmer* (*de joie*), alors que pour des constructions pronominales, seul le verbe a été isolé comme collocatif.

(4) Ses aides de camp l'entourent, empressés, respectueux, <COLLOC BASE="admiration" CAT="vpronti" FL="Sympt1" TYPESEM="manifester_physiquement">**se pâmant**</COLLOC> **d'**<LEXIQUE TYPE="affect+manif" CAT="N" DOMAINE="admiration" NV_LANGUE="courant" INTENSITE="haut" POLARITE="positif" JUGEMENT="positif" ATTITUDE="positif">**admiration**</LEXIQUE> à chacun de ses coups. (*Contes du lundi*, A. Daudet)

Outre les formes verbales composées, le collocatif peut inclure plusieurs lexies. Lorsqu'elles apparaissent obligatoires, elles ont été intégrées dans l'élément <COLLOC>. Ainsi, pour la collocation *ne pas se tenir de joie*, *ne pas se tenir* a été annoté comme collocatif, comme on peut l'observer dans l'exemple suivant.

(5) Pencroff <COLLOC BASE="joie" CAT="locvti" FL="Magn+Oper1" TYPESEM="éprouver vivement">**ne se tenait pas**</COLLOC> **de** <LEXIQUE TYPE="affect+manif" CAT="N" DOMAINE="gaieté" NV_LANGUE="courant" INTENSITE="moyen" POLARITE="positif">**joie**</LEXIQUE>, et chaque matin et chaque soir il … (*Ile mystérieuse*, J. Verne)

4.3 Adaptation des FL syntagmatiques à l'encodage des collocations

4.3.1 Les FL : un inventaire à augmenter ?

Le système des fonctions lexicales disponibles permet de rendre compte d'un nombre important de collocations dans notre tâche, et remplit ainsi bien son objectif. Cependant, quelques types de collocations récurrentes rencontrées dans les textes à annoter ne sont pas décrits par une FL standard. Pour traiter ces

cas (relativement rares), nous avons proposé, si certaines conditions étaient remplies, de dégager de nouvelles fonctions standard lors de l'encodage de relations lexicales.

Ainsi, nous avons proposé une nouvelle fonction lexicale Intent[6] (/en_vue_de/) calquée sur le modèle Propt (/à_cause_de/), qui nous paraît bien adaptée pour décrire les collocations suivantes : *pour le plaisir, pour l'amour*, par exemple :

(6) (…) quant à Jacques, trop jeune encore pour comprendre nos malheurs - il avait à peine deux ans de plus que moi -, il pleurait par besoin, <COLLOC BASE="plaisir" CAT="prep" FL="Intent" TYPESEM="en_vue_de">**pour**</COLLOC> **le** <LEXIQUE TYPE="affect" CAT="N" DOMAINE="plaisir" NV_LANGUE="courant" INTENSITE="moyen" POLARITE="positif">**plaisir**</LEXIQUE>. (*Le Petit Chose*, A. Daudet)

Même si le besoin d'instaurer la fonction Intent s'est posé d'une manière empirique et si cette fonction a été appliquée à un champ sémantique précis (lexique des émotions), nous pouvons justifier notre proposition par le fait que ce type de lien est assez systématique et qu'elle pourrait être ainsi généralisée et appliquée à d'autres lexies dans d'autres champs comme : *pour son intérêt, pour son compte, pour son bien,* etc.

4.3.2 Uniformisation du codage des FL[7]

De la même façon que pour les gloses, nous avons opté pour un encodage homogène des fonctions lexicales afin de proposer un métalangage plus simple. Nous avons essayé de réduire la diversité de FL quand cela était possible. Par exemple, à partir de deux fonctions lexicales équivalentes proposées par le Dicouèbe NonPerm1Fact0 et AntiReal1, nous avons gardé une seule valeur, AntiReal1, qui apparaît plus facilement décodable, par exemple dans :

Surmonter l'angoisse: NonPerm1Fact0 - /[X] ne pas se laisser influencer par son A./
Surmonter la crainte : AntiReal1 - /[X] ne pas se laisser influencer par sa C./

Dans certains cas où les FL s'avéraient difficiles à coder, nous avons eu recours au Dicouèbe. Cependant, le traitement proposé était assez complexe et parfois non uniforme, ce qui semble montrer la difficulté d'utilisation de ce métalangage. Ainsi, par exemple pour *épargner la déception* / *épargner la peine*, le dictionnaire propose deux FL différentes :

Epargner la déception : NonPermOper21 / [Qqch./Qqn.] empêcher que X éprouve une D./
Epargner la peine : CausNonIncepFunc1 / [Qqch.] empêcher que X éprouve de la P./

Pour uniformiser le codage, nous proposons d'encoder dans les deux cas la fonction NonPermFunc1 pour les exemples comme : *épargner*, *éviter, empêcher + Naffect*.
En bref, pour certaines fonctions complexes, il nous semble que le modèle gagnerait peut-être à proposer un encodage plus systématique.

4.3.3 Fonction lexicale syntagmatique ou paradigmatique ?

Une autre difficulté d'adaptation du modèle des FL a été le recours à certaines fonctions paradigmatiques pour exprimer des relations de cooccurrences (Cf. aussi Alonso Ramos & Tutin 1996 sur ce point). Certaines fonctions comme A_1 sont en effet essentiellement décrites comme ayant un fonctionnement paradigmatique :

A_1 : adjectif typique pour le premier actant du mot clé. Exemple : A_1(bonheur) = heureux.

[6] Du lat. *intentio,* étiquette proposée afin de suivre les appellations latines du DEC.
[7] Nous tenons à remercier Alain Polguère pour ses suggestions et commentaires sur le traitement.

Il arrive cependant très fréquemment dans le champ des émotions que des collocations qui incluent le mot clé puissent être aussi décrites à l'aide de cette fonction standard : *en colère, en joie, en amour* (fr. québécois), *dans le désespoir*. Nous préférerions donc dans ce cas recourir à une fonction qui indique une relation syntagmatique, plutôt qu'utiliser de façon détournée une FL paradigmatique[8]. Une solution serait de proposer une nouvelle fonction syntagmatique qui indique la préposition typique pour l'état du premier actant du mot-clé. On aurait ainsi :

Loc_1 : préposition qui décrit l'état pour l'actant 1 du mot-clé.
Loc_1(désespoir) = *dans le* ~
Loc_1(colère) = *en* ~

Cependant, pour éviter la prolifération des FL et conserver une cohérence avec le traitement du DEC qui reste notre base, nous avons conservé les FL A_1 et Adv_1, tout en proposant des gloses spécifiques. Du point de vue de la cohérence du modèle cependant, cela ne nous paraît pas parfaitement satisfaisant.

4.4.4 Le traitement des metaphores

En travaillant sur le lexique abstrait que sont les émotions, nous avons été souvent confrontées au problème de la valeur figurée de certaines occurrences. En effet, certains collocatifs des noms d'affect pourraient être qualifiés de métaphoriques. C'est en particulier le cas des collocatifs véhiculant une valeur intensive, comme des nominaux : *éclair de joie, feu de l'amour, transport de fureur* ou les collocations verbales : *plonger dans la tristesse, se noyer dans le chagrin* ou dénotant une valeur aspectuelle : *refroidir l'enthousiasme*.

Dans la liste des fonctions lexicales proposée par le DEC, on relève une fonction Figur qui renvoie à une « métaphore codifiée par la langue dont la combinaison avec le mot clé est un synonyme (plus étroit) du mot clé » (Mel'čuk et al., 1984:7). Cette fonction lexicale est attribuée par exemple à quelques substantifs, comme une fonction lexicale simple ou dans les configurations avec d'autres fonctions paradigmatiques ou syntagmatiques :

Figur (haine) = feu [de la]	/Métaphore/
MultFigur (vapeur) = nuage [de]	/Une certaine quantité de ~/
AntiMagn.Figur (espoir) = lueur [d']	/Métaphore d'un ~ peu intense/

Nous pouvons observer que cette fonction n'est pas attribuée d'une manière systématique. Par exemple, *débordement (d'enthousiasme)* est codé dans le Dicouèbe par Magn+Figur, mais *déborder (d'enthousiasme)* est codé Magn+Oper1 sans mention de l'aspect métaphorique. Ainsi, il y a parfois des variabilités dans le codage de ce type de collocations et par la suite aussi de leur paraphrasage.

Nous trouvons cette fonction assez difficile à manipuler dans le corpus. D'après la définition de Figur dans le DEC, il s'agit d'une des fonctions paradigmatiques qui ne modalisent pas les collocations mais les rapports sémantiques entre les éléments, notamment dans le cas de Figur, une relation de synonymie. Certains précisent qu'il s'agit d'un « synonyme plus riche » de la base, ce qui implique qu'il rajoute une valeur sémantique supplémentaire. Cette définition en termes de synonymie apparaît discutable, ainsi que le statut paradigmatique de cette fonction. En effet, le collocatif figuré instaure une relation syntagmatique avec la base et cette relation devrait être encodée systématiquement. La fonction Figur dénote le plus souvent le haut degré d'intensité et c'est peut-être pour cela qu'elle est parfois identifiée aux autres FL sans indiquer la valeur sémantique véhiculée (p.ex. dans : Figur (haine) = feu [de la]).

Nous proposons de garder la fonction Figur, mais en la surajoutant à une description au niveau syntagmatique. Il faudrait l'appliquer à tous les cas des collocations à valeur figurée. Par exemple :

[8] De la même façon qu'on distingue dans le DEC V_0 et $Oper_1$.

MagnOper1+Figur (amour) = brûler [d']
CausPredMinus+Figur (enthousiasme) = refroidir [ART ~]
S1Magn+Figur (haine) = feu [de la]

Il serait aussi préférable de coder Figur d'une autre manière, par exemple avec le pointeur ou entre parenthèses, afin de mieux souligner qu'il s'agit d'une information d'un autre niveau, superposée à la description de collocation avec les FL.

5 Bilan et conclusion

Au terme de cette expérimentation, un bilan s'impose. Du point de vue quantitatif, 1892 collocations mettant en jeu des mots d'émotions ont été codées. 93,9% d'entre elles ont pu être traitées à l'aide fonctions lexicales standard. Seulement 4,4% de fonctions lexicales annotées n'ont pas reçu de glose sémantique, comme on peut l'observer dans le tableau 4. Cela montre que le système des FL standard propose une réponse satisfaisante pour l'encodage de la majorité des collocations.

Type de FL	**Proportion dans le corpus**
FL standard	93,9% (=1776/1892)
FL non standard	6,1% (=116/1892)
FL standard avec glose	96,6% (=1716/1776)
FL standard sans glose	4,4% (=60/1776)

Tableau 4. Proportions de FL annotées

Si l'on se tourne maintenant vers les FL les plus utilisées (Cf. tableau 5), on retrouve bien les FL souvent citées dans la littérature sur les collocations : les verbes supports (Oper1) représentent à eux seuls un quart des collocations annotées et la collocation Magn (surtout quand elle est adjectivale) est également très courante. Les *blessé grave, faim de loup* et *peur bleue* souvent cités correspondent donc bien à un prototype productif. Les difficultés d'encodage ne concernent donc qu'un petit nombre de collocations, ce qui en minimise la portée.

Fonction Lexicale	**Nombre total**	**Pourcent**
Oper1	442	24,8%
Magn (adj)	295	16,6%
CausFunc(0/1)	205	11,5%
Caus(1)Manif	68	3,8%
Magn (adv)	54	3%
S0Sympt1	53	2,9%
Sympt1	51	2,9%
S0Manif1	44	2,4%

Tableau 5. Répartition des principales FL standard

D'une manière générale, notre expérimentation montre que le système des fonctions lexicales permet de coder la plupart des collocations de façon satisfaisante, même si quelques points gagneraient à être traités pour garantir la cohérence du système. Une simplification et une homogénéisation du modèle en permettraient probablement une meilleure utilisation.

Enfin, il nous reste maintenant à tester l'exploitation de ces ressources annotées pour les applications didactiques que nous visons.

Remerciements

Nous remercions tout particulièrement Alain Polguère, qui nous a fourni de précieuses explications sur le codage des Fonctions Lexicales. Un grand merci également à Thierry Fontenelle qui nous a autorisées à utiliser sa BD de

collocations, extraite du dictionnaire Collins-Robert. Hormis les auteures de ces lignes, le corpus a aussi été annoté par Gwendoline Bloquet et Mériam Haddara que nous remercions aussi vivement. Merci aussi à nos collègues Cristelle Cavalla et Francis Grossmann pour leurs remarques avisées.

References

Alonso Ramos, Margarita, & Agnès Tutin. 1996. A Classification and description of the Lexical Functions of the Explanatory Combinatorial Dictionary for the treatment of LF Combinations. In Wanner L. (ed.), *Lexical Functions in Natural Language Processing and Lexicography*. John Benjamins, Amsterdam, 146-167.

Augustyn, Magdalena, Sabrina Ben Hamou, Gwendoline Bloquet, Vannina Goossens, Fanny Rinck, Mathieu Loiseau. 2008. Constitution de ressources pédagogiques numériques : le lexique des affects. *Autour de la langue et du langage : perspective pluridisciplinaire*. Presses Universitaires de Grenoble, Grenoble.

Cavalla, Cristelle, & Virginie Labre. A paraître. L'enseignement en FLE de la phraséologie du lexique des affects. In Novakova I. & Tutin A. *Lexique des émotions*. Ellug, Grenoble.

Charest, Simon, Eric Brunelle, Jean Fontaine, Bertrand Pelletier. 2007. Élaboration automatique d'un dictionnaire de cooccurrences grand public. *Actes de Traitement Automatique des Langues Naturelles 2007*, Toulouse, 282-292.

Goossens, Vannina. 2005. Les noms de sentiment : esquisse de typologie sémantique fondée sur les collocations verbales. *Lidil*, 32:103-121.

Fellbaum, Christiane, & Alexander Geyken. 2005. Transforming a Corpus into a Lexical Resource: The Berlin Idiom Project. *Revue Française de Linguistique Appliquée*, X (2):45-62.

Fontenelle, Thierry. 1997. *Turning a Bilingual Dictionary into a Lexical-Semantic Database*. (Lexicographica/Series maior). Niemeyer Verlag, Tübingen.

Granger, Sylviane. 1998. Prefabricated patterns in advanced EFL writing: collocations and formulae. Cowie A. (ed.) *Phraseology: theory, analysis and applications*. Oxford University Press, Oxford, 145-160.

Hausmann, Franz Josef. 1989. Le dictionnaire de collocations. Hausmann, F.J., Reichmann, O., Wiegand, H.E., Zgusta, L. (eds), *Wörterbücher : ein internationales Handbuch zur Lexicographie. Dictionaries. Dictionnaires*. De Gruyter, Berlin/New-York, 1010-1019.

Ludewig, Petra. 2001. LogoTax : un outillage exploratoire pour l'étude de collocations en corpus. *Traitement automatique du langage*, 42(2):623-642.

Mel'čuk, Igor A., & Alain Polguère. 2007. *Lexique actif du français. L'apprentissage du vocabulaire fondé sur 20 000 dérivations sémantiques et collocations du français*. De Boeck Duculot, Louvain-la-Neuve.

Mel'čuk, Igor A., André Clas, Alain Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*. Duculot, Paris.

Mel'čuk, Igor A. et al. 1984, 1988, 1992, 1999. *Dictionnaire explicatif et combinatoire du français contemporain*, Vol. 1, 2, 3, 4. Presses de l'Université de Montréal, Montréal.

Nesselhauf, Nadja. 2005. Collocations in a Learner Corpus. *Studies in Corpus Linguistics*, 14. John Benjamins, Amsterdam.

Rinck, Fanny, & Agnès Tutin. 2007. Annoter la polyphonie dans les textes : le cas des passages entre guillemets. *Corpus*, 6:79-100.

Silberztein, Max. 1999. INTEX: a Finite State Transducer toolbox. *Theoretical Computer Science*, 231-1:33-46. Elsevier Science, Saint-Louis.

Tréville, Marie-Claude, Lise Duquette. 1996. *Enseigner le vocabulaire en classe de langue*. Hachette, Paris.

Tutin, Agnès. 2004. Pour une modélisation dynamique des collocations dans les textes. *Actes d'Euralex*. Lorient, 6-10 juillet 2004.

Tutin, Agnès. 2005. Annotating Lexical Functions in Corpora : Showing Collocations in Context. 2nd International Conference on the Meaning-Text Theory. Moscow, June 23-25 2005.

Structuration et balisage sémantique des définitions du *Trésor de Langue Française informatisé* (TLFi)

Lucie Barque & Alain Polguère
OLST—Université de Montréal
C.P. 6128, succ. Centre-ville,
Montréal (QC) H3C 3J7 — Canada
lucie.barque, alain.polguere@gmail.com

Abstract

We present an ongoing lexicographic project which aims at providing an explicit and formalized structuring for the definitions of the French electronic dictionary *Trésor de la Langue Française informatisé* (TLFi). Section 1 deals with the general issue of formal structuring of lexicographic definitions; section 2 introduces the notion of analytical definition and gives an overview of its use in French lexicography, particularly in the TLFi; section 3 details our project, its methodology and consecutive steps.

1 Problème de la structuration des définitions lexicographiques

Il n'existe pas à l'heure actuelle pour le français (et, à notre connaissance, pour aucune autre langue) de base de données lexicale, libre d'accès et à très large couverture, proposant pour chaque unité lexicale décrite une définition explicitement structurée. Par *définition explicitement structurée*, nous entendons une définition lexicographique munie d'un balisage formel (XML ou autre) indiquant :

1. la structure en composantes définitionnelles de type genre prochain et différences spécifiques ;
2. le rôle joué par chaque composante dans la caractérisation du sens de l'unité lexicale définie.

Examinons, à fin d'illustration, la définition (1) ci-dessous de l'acception de base de TROTTINETTE, tirée du *Trésor de la Langue Française informatisé*, désormais *TLFi* (Dendien & Pierrel, 2003).

(1) TROTTINETTE (acception **A.1**) : Jouet composé d'une plate-forme allongée montée sur deux petites roues et d'un guidon à direction articulée, que l'enfant fait avancer en s'aidant d'un pied qu'il pose régulièrement par terre pour donner l'impulsion ou en actionnant une pédale en un mouvement de va-et-vient.

Une version structurée de cette définition devrait indiquer explicitement :
– la présence de trois composantes définitionnelles majeures ;
– le fait que *jouet* est la composante centrale — identifiant le genre prochain —, qui caractérise l'unité lexicale décrite comme dénotant un type de 'jouet' ;
– le fait que les composantes (i) *composée … direction articulée* et (ii) *que l'enfant … va-et-vient* sont des composantes périphériques — différences spécifiques —, qui spécifient la composante centrale, respectivement, en termes de 'parties constituantes' et 'mode d'utilisation'.

Il est clair qu'une base de données de définitions lexicographiques structurées d'une telle façon serait extrêmement précieuse, aussi bien pour la communauté du traitement automatique de la langue (Barnbrook, 2002, pages 2–7), que pour la recherche en linguistique et la lexicographie.

Dans le présent article, nous présentons un projet, en cours d'exécution, qui vise la conversion des définitions du TLFi sous une forme structurée. Ce projet s'appuie sur les acquis du projet BDéf de base de définitions lexicales formalisées (Altman & Polguère, 2003; Barque, 2008). Contrairement à ce qui a été fait avec la BDéf, nous visons une formalisation se limitant à la caractérisation de la structure des définitions en termes de blocs définitionnels, alors que la DBéf s'attachait aussi à la formalisation des composantes elles-mêmes en termes de propositions élémentaires écrites dans un langage contrôlé. Cependant, le présent projet est plus ambitieux dans la mesure où il vise le développement à moyen terme d'une ressource lexicale à très large couverture, alors que le travail sur la BDéf s'est concentré sur la formalisation d'un sous-ensemble du lexique français : celui décrit dans les quatre volumes publiés du *Dictionnaire explicatif et combinatoire du français contemporain* ou DEC (Mel'čuk et al., 1984, 1988, 1992, 1999).

Il y a deux raisons pour lesquelles nous appuyons notre travail sur le TLFi, plutôt que sur tout autre dictionnaire du français. Premièrement, les définitions de ce dictionnaire sont dans leur grande majorité des définitions, appelées *définitions analytiques*, qui respectent certaines contraintes de forme et de contenu permettant leur structuration formelle en termes de composantes définitionnelles [1]. Deuxièmement, ce dictionnaire, développé durant plus de trente ans à l'INALF (maintenant ATILF), est la seule modélisation complète de haute qualité du lexique français librement accessible sous forme électronique à fin de recherche [2].

Dans ce qui suit, section 2, nous spécifierons ce qu'est une définition analytique, puisque c'est sur ce type particulier de définition lexicographique que nous travaillons. Ensuite, section 3, nous décrirons le projet actuellement en cours : corpus traités, stratégie adoptée et principales étapes de traitement des données.

2 Notion de définition analytique

2.1 Délimitation de la notion

Il existe une ample littérature sur la définition lexicographique, littérature concernant aussi bien les aspects strictement lexicographiques de la question — par exemple, (Wierzbicka, 1987; Dostie et al., 1999; Barnbrook, 2002; Rundell, 2008) — que ses aspects plus théoriques — par exemple, (Fodor et al., 1980; Wierzbicka, 1996; Riemer, 2006). Il ne serait pas pertinent de faire ici une synthèse de tout ce qui a été dit sur cette notion et nous nous contenterons, tout d'abord, d'une caractérisation approximative de ce que l'on peut entendre par *définition lexicographique*.

> Une définition lexicographique d'une lexie [3] L est (i) une analyse de son sens, (ii) qui prend la forme d'une paraphrase linguistique de L.

Bien qu'assez vague, cette caractérisation exclut du domaine de la définition lexicographique cetaines modélisations sémantiques des lexies, comme par exemple celles du *Lexique génératif* (Pustejovsky, 1995), ou toutes autres modélisations fondées sur le langage de la logique formelle, puisqu'il ne s'agit pas à proprement parler d'**énoncés linguistiques** paraphrasant les lexies décrites.

La caractérisation ci-dessus exclut aussi les descriptions qui ne seraient pas des **analyses** sémantiques. Ainsi, Atkins & Rundell (2008, pages 36–38) considèrent que, parmi les trois descriptions sémantiques de l'acception de base du vocable anglais DISTURB données ci-dessous, seules les deux dernières ont le statut de paraphrases définitionnelles :

(2) a. to intrude on ; interrupt [*Collins English Dictionary*, 2006]

b. If you disturb someone, you break their rest, peace or privacy [*Collins School Dictionary*, 2006]

c. To interrupt someone and stop them from continuing what they were doing [*Macmillan English Dictionary for Advanced Learners*, 2002]

1. Nous traiterons brièvement des cas de définitions non analytiques à la fin de la section suivante.

2. Nous sommes infiniment reconnaissants à Jean-Marie Pierrel, directeur de l'ATLIF, de nous avoir donné accès aux fichiers source du TLFi.

3. C'est-à-dire, une unité lexicale. Dans cet article, nous utilisons la terminologie de l'approche de la Lexicologie Explicative et Combinatoire, telle que présentée dans (Mel'čuk et al., 1995) ou (Polguère, 2008).

Examinons brièvement les caractéristiques propres à chacune de ces descriptions :

– Contrairement à ce que disent Atkins & Rundell, la description (2a) a recours au paraphrasage, puisque le quasi-synonyme est la plus élémentaire des paraphrases. Ils ont cependant raison de considérer qu'une telle description ne **définit** pas le sens, car elle ne l'analyse pas. L'équivalent quasi-synoymique n'est donc pas à proprement parler une définition lexicographique (une analyse du sens) [4].
– La description (2b), quant à elle, définit véritablement DISTURB, en mettant au jour son contenu sémantique. On pourrait se questionner quant à la nature paraphrastique de la description proposée. Nous sommes cependant d'accord pour considérer avec Atkins & Rundell qu'il s'agit bien là d'un paraphrasage. En effet, l'énoncé linguistique *If you disturb someone...* peut être facilement reformulé sous forme d'une égalité du type « definiendum = definiens » :

  X disturb Y = X break Y's rest, peace or privacy

  Le format choisi par le dictionnaire est simplement considéré par les rédacteurs comme plus approprié, d'un point de vue pédagogique, que celui que nous allons maintenant examiner.
– La description (2c) est bien entendu elle aussi une analyse paraphrastique ; contrairement à la précédente, elle ne prend en compte que le définiens. Elle semble plus rigoureuse, plus « académique » d'un certain point de vue, mais on peut noter qu'elle présente l'inconvénient de ne pas introduire de façon explicite la structure actancielle complète de la lexie définie et sa correspondance dans la paraphrase, contrairement à (2b), qui a recours aux deux pseudo-variables *you* and *someone* pour nommer, respectivement, le premier et le second actant de DISTURB [5].

Nous voyons donc que le critère de paraphrase (pris dans un sens large) n'est pas suffisant pour caractériser un type de définition considéré comme satisfaisant : il faut aussi que la définition analyse le sens de l'unité lexicale définie en termes de sens plus simples. Un tel principe est bien entendu appliqué dans nombre de dictionnaires grand public, ainsi qu'en lexicographie théorique — notamment, en Lexicographie Explicative et Combinatoire (Mel'čuk et al., 1995) et dans l'approche du *Natural Semantic Metalanguage* (Wierzbicka, 1996).

La dernière contrainte couramment appliquée aux définitions lexicographiques qui vont nous intéresser ici — et qui n'est pas présente dans la caractérisation des définitions lexicales donnée plus haut – est le principe aristotélicien de décomposition par genre prochain et différences spécifiques : on doit retrouver au cœur de la définition une composante centrale (= genre prochain) qui correspond au noyau sémantique hyperonymique de l'unité définie, accompagnée d'une ou plusieurs composantes périphériques (= différences spécifiques) qui particularisent l'unité définie par rapport à son genre prochain et à ses cohyponymes. Sémantiquement, la composante centrale de la définition est la composante sémantiquement dominante de la représentation sémantique de celle-ci [6] ; elle a de ce fait une fonction classifiante, dans la mesure où elle permet de regrouper des lexies ayant un même noyau sémantique. Syntaxiquement, il s'agit d'un lexème ou d'un syntagme qui est au sommet de la structure syntaxique de la définition. Ainsi, *to interrupt* est la composante centrale de la définition (2b) ci-dessus.

Les définitions lexicographiques qui respectent tous les critères qui viennent d'être examinés — analyses paraphrastiques formulées en termes de genre prochain et différences spécifiques — sont appelées *définitions analytiques* dans (Polguère, 2008, page 183). C'est désormais spécifiquement à ce type de définitions que nous nous intéresserons dans la suite de cet article.

4. On notera que ce point de vue n'est pas nécessairement partagé par tous. Ainsi, Benson et al. (1986, pages 203–204) considèrent le recours aux (quasi-)équivalents synonymiques comme un type particulier de définition lexicographique, qu'ils appellent *synonym definition*.

5. Voir (Rundell, 2008), pour une brève comparaison entre les définitions sous forme de phrase complète (angl. *full-sentence definitions*), du type (2b), et les définitions strictement paraphrastiques, du type (2c).

6. Sur la dominance communicative dans les réseaux sémantiques, voir (Polguère, 1997).

2.2 Pratique lexicographique

Le recours à la définition analytique est la norme dans les dictionnaires de référence du français tels le TLFi [7] ou le *Nouveau Petit Robert*. On doit cependant souligner trois cas fréquents de transgression de ce principe définitionnel dans ces dictionnaires.

Premièrement, il est bien connu que les dictionnaires décrivent le contenu sémantique et/ou syntaxique des lexies grammaticales (articles, prépositions régies du type À et DE, etc.) par des énoncés qui ne sont en aucun cas des paraphrases analytiques, comme l'illustrent les deux descriptions ci-dessous de l'article indéfini français, tel qu'employé dans *Marc veut avoir **une** petite sœur* :

(3) a. UN (acception **II.A.1**) : Désigne un objet, un élément distinct mais indéterminé. [*Petit Robert*, 2007]

b. UN^{2} (acception **I.A.1**) : [Empl. spécifique. *J'ai acheté un livre pour enfants.* Ce qui est dit est vrai d'un seul livre, pris sur l'ensemble des livres pour enfants. Certes, p. oppos. à *le livre*, ce livre n'appartient pas encore au thème de l'énoncé ; il est indéterminé, mais il s'agit d'un livre précis] [TLFi]

Dans le cas de la description (3b), l'usage des crochets indique une caractérisation « sémantico-grammaticale » clairement distinguée d'une définition lexicographique. Il arrive cependant que, dans le TLFi, les caractérisations de ce type et les définitions non analytiques cohabitent pour un même article de lexie, comme l'illustre la caractérisation ci-dessous de la préposition DE, telle qu'employée dans *naviguer **de** Brest à Halifax* :

(4) DE^{2} (acception **I.A.1**) : [Orig. spatio-temporelle.] Le point de départ se situe dans l'espace ou dans le temps (s'oppose à la prép. *à*, parfois à *en*, plus rarement à *jusqu'à*), par rapport à un point d'aboutissement dans l'espace ou dans le temps. [TLFi]

Le deuxième cas de non respect du principe de définition analytique est tout simplement le recours à des descriptions par quasi-synonymes, qui ne sont pas, comme nous l'avons vu en 2.1, des analyses du sens. Le TLFi utilise fréquemment les « définitions » par synonymes, comme illustré sous (5) :

(5) FUNAMBULESQUE (acception « figurée ») : Fantaisiste, bizarre.

On trouve aussi dans le TLFi beaucoup les reformulations successives d'un même sens ou de sens très proches, comme illustré sous (6) :

(6) a. FRANC^{3} (acception **I.A.3**) : Qui est moralement libre ; qui agit de sa propre volonté.

b. EN ÊTRE POUR SES FRAIS (acception « figurée » de la locution, sous FRAIS^{2} **C.3**) : S'être mis inutilement en peine, être déçu.

c. FUSTIGER (acception **B**) : Attaquer, combattre, critiquer violemment.

Ce type de définitions pose problème car il faut pouvoir distinguer les cas où il s'agit vraiment d'une reformulation (FRANC^{3} **I.A.3**), des cas où il s'agit d'une unité vague, c'est-à-dire d'une unité qui peut dénoter l'un et/ou l'autre des deux sens figurant dans la paraphrase (EN ÊTRE POUR SES FRAIS). Les reformulations peuvent en outre donner lieu à des ambiguïtés de rattachement concernant certaines composantes. On est par exemple en droit de se demander si la composante *violemment*, dans la définition de FUSTIGER **B**, est à rattacher seulement à *critiquer* ou à la composante plus large *attaquer, combattre, critiquer*.

Finalement, le dernier cas de non respect du principe de définition analytique est plus difficile à identifier. Il concerne les définitions paraphrastiques qui, de par leur structure syntaxique, présentent comme genre prochain une composante qui devrait normalement se retrouver parmi les différences spécifiques. Nous avons rencontré un tel cas avec la définition de TROTTINETTE du TLFi — voir (1), en début d'article. La structure syntaxique de cette définition met en évidence le fait que TROTTINETTE est classifiée sémantiquement

7. Voir (Frassi, 2007) pour une étude détaillée des définitions du TLF.

dans le TLFi en tant que ‘jouet’, et non en tant que ‘véhicule (à deux roues)’. Le fait qu’une trottinette est avant tout un véhicule — et seulement secondairement un jouet — n’est qu’implicitement suggéré dans la définition, à travers la description du mode d’utilisation. Au moment de la rédaction de cette définition, les trottinettes étaient utilisées comme jouets et n’avaient pas encore connu leur seconde vie (sous une forme « relookée ») en tant que mode de déplacement urbain branché. Cependant, depuis toujours, la trottinette est avant tout un type de véhicule (‘artefact servant à se déplacer’) et n’est un jouet que de façon secondaire : une trottinette que l’on utilise pour se rendre au travail, et non pour jouer, reste une trottinette.

Ici s’achève cette section sur la notion de définition analytique et son implication dans la structuration **implicite** des définitions des dictionnaires de langue. Nous allons maintenant aborder la présentation de notre projet de structuration formelle explicite de ce type de définitions, projet fondé sur l’utilisation du TLFi en tant que ressource lexicale.

3 Vers une base de définitions formalisées dérivée du TLFi

Nous abordons maintenant la présentation de notre projet de structuration des définitions du TLFi. Après avoir donné quelques informations sur la taille du corpus traité (section 3.1), nous présenterons la stratégie générale adoptée pour ce projet (section 3.2). Puis, nous décrirons les trois étapes principales de sa réalisation : mise en évidence, au moyen d’un balisage XML, de la structuration des définitions en composantes centrale et périphériques (section 3.3), marquage sémantique normalisé de ces composantes (section 3.4) et mise en œuvre d’une base lexicale sémantique dérivées du TLFi (section 3.5).

3.1 Corpus traité

L’identification de la nomenclature de toutes les entités lexicales définies du TLFi n’est pas aisée à faire. En effet, le TLFi compte 54 281 entrées principales (vocables potentiellement polysémiques), auxquelles s’ajoutent 18 096 entrées « enchâssées » correspondant à des dérivés morphologiques d’une entrée principale (par exemple, FRUITARISME sous FRUIT) [8]. Le TLFi contient aussi 59 168 syntagmes définis, syntagmes qui peuvent être de deux types. Tout d’abord, le syntagme défini peut être une locution : par exemple, FRUIT DÉFENDU sous FRUIT**[1]** ; il s’agit là d’une pratique généralisée dans les dictionnaires de langue [9]. De façon plus originale, nombre de collocations sont aussi définies comme des touts lexicaux, soit dans l’article de la base de la collocation (7a), soit dans celle de son collocatif (7b).

(7) a. *Vol domestique* (sous VOL**[2] A.1**) : Vol commis par un domestique, un employé de maison, soit envers son employeur à l’intérieur des locaux que celui-ci possède et dans lesquels il l’accompagne, soit envers des personnes se trouvant dans l’habitation de son employeur.

b. *Troupes fraîches, chevaux frais* (sous FRAIS**[1] C.1b**) : Troupes, chevaux destinés à remplacer ceux qui sont fatigués.

Pour notre projet, nous avons extrait automatiquement les 271 164 définitions décrivant les sens de ces différents types d’entités lexicales [10].

3.2 Stratégie adoptée

Notre stratégie de développement de la base de définitions est avant tout lexicographique. Nous entendons par là qu’elle s’appuie en premier lieu sur un travail manuel effectué par des lexicographes et des annotateurs [11]. On peut mettre en opposition notre approche d’extraction de données lexicographiques struc-

8. Un peu comme le dictionnaire *Lexis* de Larousse, le TLF subordonne la description de nombreux dérivés morphologiques à celle de leur source morphologique. Il semblerait toutefois que seuls les dérivés rares ou techniques soient traités de cette façon.

9. Même s’il est clair que chaque locution devrait en théorie posséder son entrée propre dans la nomenclature, comme cela est le cas dans le DEC (Mel’čuk et al., 1984, 1988, 1992, 1999).

10. Cette extraction est aisée dans la mesure où les fichiers informatiques du TLFi contiennent un balisage structurant les articles lexicographiques ; ainsi, toutes les définitions sont encadrées par les deux balises `<def>` et `</def>`.

11. Voir, à ce propos, notre guide d’annotation (Barque & Polguère, 2009).

turées avec celle adoptée dans (Barnbrook, 2002) pour l'analyse des définitions du *Collins Cobuild Student's Dictionary* (Sinclair, 1990), qui repose sur des procédures entièrement automatiques. Notre stratégie nous semble plus appropriée pour le présent projet dans la mesure où nos premières observations nous conduisent à penser que l'ensemble des définitions du TLFi forme un tout moins uniforme que l'ensemble des défnitions de dictionnaires commerciaux tels que le *Collins Cobuild* ou, mieux, le *Longman Dictionary of Contemporary English* LDOCE (cf. son métalangage définitionnel contrôlé) [12].

Toutefois, nous n'écartons pas la voie de l'automatisation du traitement, mais en tant que support à l'analyse manuelle. Le corpus à traiter étant énorme (voir *supra*), nous testons actuellement une pré-annotation automatique au moyen du logiciel MACAON, développé par Alexis Nasr [13]. Ce prétraitement présente en outre un intérêt lexicographique, puisque MACAON nous permet de décrire la grammaire des définitions du TLFi et de voir ainsi dans quelle mesure le métalangage définitionnel du TLFi se rapproche d'un sous-langage (Kittredge, 1982) et dans quelle mesure il pourrait être amélioré (voir plus bas, section 3.5).

3.3 Étape 1 : structuration des définitions

La première étape de notre projet, en cours de réalisation, consiste à structurer les définitions du TLFi. Plus précisément, il s'agit de délimiter la composante centrale et les éventuelles composantes périphériques de la paraphrase définitionnelle. Considérons, pour commencer, la définition non structurée de BLINDÉ$_N$:

(8) BLINDÉ$_N$ (nom décrit à l'intérieur de l'article du participe BLINDÉ **II.A**) : Un véhicule militaire blindé.

La tâche la plus délicate consiste à délimiter la composante centrale, qui doit être, rappelons-le, une composante sémantiquement classifiante. Dans notre exemple, il est possible d'hésiter entre *véhicule*, *véhicule militaire* ou, même, *véhicule militaire blindé*, puisque chacune de ces expressions peut se retrouver comme composante centrale de la définition de plusieurs lexies françaises. En l'occurrence, dans le TLFi, *véhicule militaire blindé* ne se retrouve que dans la définition de BLINDÉ$_N$ **II.A** ; mais le contenu sémantique de cette expression se retrouve, exprimé différemment, dans la définition d'autres lexies, comme CHAR :

(9) CHAR (acception **C.3**) : **Engin de guerre motorisé et blindé**, monté sur chenilles et doté d'un armement (mitrailleuses, canons, etc.) et que manœuvrent des soldats placés à l'intérieur.

On choisira donc ici le syntagme le plus long comme composante centrale, sans composante périphérique. Voyons maintenant une définition dont la structuration sera plus standard, celle de BROUETTE :

(10) BROUETTE (acception **B.1**) : Véhicule à une roue et à deux brancards servant au transport des matériaux.

On peut se poser ici le même type de question que dans le cas précédent : faut-il choisir *véhicule* ou *véhicule à une roue* comme composante centrale ? Dans ce cas-ci, la composante centrale doit être *véhicule* car le TLFi ne contient pas de définition d'autres lexies dénotant des véhicules à une roue [14]. Une fois la composante centrale (*véhicule*) délimitée, il devient relativement simple d'isoler les différentes composantes périphériques. Nous présentons ci-dessous la version structurée de cette définition : la composante centrale y est indiquée au moyen d'une balise <CC> et les composantes périphériques au moyen de balises <CP>.

(11) BROUETTE (acception **B.1**)

```
<PARAPH>
  <CC>Véhicule</CC>
  <CP>à une roue et à deux brancards</CP>
  <CP>servant au transport des matériaux</CP>
</PARAPH>
```

12. À propos du traitement automatique des définitions du LDOCE, voir (Fontenelle, 2009), qui vient de nous être signalé et que nous n'avons malheureusement pas encore eu la possibilité de consulter.

13. http://pageperso.lif.univ-mrs.fr/~alexis.nasr/macaon/index.html

14. Il existe bien une entrée SIDE-CAR. Cependant, l'acception de base dénote une ≪ **Caisse** carossée monoplace à une roue ≫. Seule la seconde acception, métonymique, dénote un ≪ **Véhicule** formé par la réunion d'une motocyclette et de cette caisse ≫.

On remarque ici que la séquence *à une roue et à deux brancards* ne forme pas deux mais une seule composante périphérique. Les deux éléments de la conjonction participent en effet de la même façon à la spécification du sens : il s'agit d'indiquer des parties caractéristiques du véhicule. Comme on le voit, nous procédons à une segmentation minimale en composantes périphériques. La règle que nous appliquons consiste à ne délimiter, dans la mesure du possible, que des composantes périphériques ayant un apport de nature distincte vis-à-vis de la composante centrale.

3.4 Étape 2 : Marquage sémantique et construction d'une hiérarchie d'étiquettes sémantiques

Une fois identifiée la structure générale de la définition, nous procéderons au marquage sémantique de ses composantes. Le marquage sémantique consiste tout d'abord à attribuer à la composante centrale une **étiquette sémantique**, c'est-à-dire une expression linguistique normalisée qui rend compte de la valeur sémantique de la composante centrale (Polguère, 2003). Pour ce faire, nous extrairons de notre base de définitions segmentées des ensembles de définitions qui ont une composante centrale identique ou proche, par exemple la composante *véhicule* et toute autre composante elle-même définie par *véhicule* (*voiture*, etc). Puis, dans le cadre du développement de la hiérarchie des étiquettes sémantiques du TLFi [15], nous créerons le jeu d'étiquettes sémantiques qui servira à marquer ces différentes composantes centrales. Dans la série de définitions annotées présentée ci-dessous, les deux étiquettes servant à marquer les composantes centrales sont `véhicule` et `voiture` (étiquette fille de `véhicule`, dans la hiérarchie).

(12) a. BROUETTE : `<CC etiq=`**`véhicule`**`>Véhicule</CC> ...`

b. CARAVANE : `<CC etiq=`**`véhicule`**`>Roulotte</CC> ...`

c. BERLINE : `<CC etiq=`**`voiture`**`>Voiture automobile</CC> ...`

d. ACCÉLÉRIFÈRE : `<CC etiq=`**`voiture`**`>Sorte de diligence ou voiture publique</CC> ...`

Notons que les étiquettes sémantiques doivent être désambiguïsées par le numéro d'acception qui leur correspond dans le TLFi et que leur sens doit être décrit par la définition correspondant à cette acception, comme illustré sous (13a-b).

(13) a. VÉHICULE (acception **II.A**) : Engin constitué d'un châssis muni de roues, à traction animale ou autopropulsé, servant au transport routier ou ferroviaire.

b. VOITURE (acception **B.2**) : Véhicule automobile servant à transporter un nombre réduit de personnes ou des objets de faible encombrement.

Une fois la composante centrale marquée d'une étiquette, il reste à indiquer le **rôle** que joue chaque composante périphérique par rapport à cette composante centrale [16]. La définition de l'étiquette sémantique nous permet déjà de proposer une première série de rôles de composantes périphériques. Par exemple, la définition de l'étiquette `véhicule`, donnée ci-dessus en (13a), présente les **parties** caractéristiques de l'engin (*constitué d'un châssis muni de roues*), son **mode de fonctionnement** (*traction animale ou autopropulsé*) et, enfin, sa **fonction** (*servant au transport routier ou ferrovière*). L'observation des définitions des lexies étiquetées `véhicule` nous permet de confirmer ces différents rôles de composantes périphériques, comme on le voit dans les définitions de REMORQUE (14) et LANDAU (15), et, le cas échant, d'en proposer d'autres : les définitions de BOLIDE (16) et TACOT (17) nous indiquent, par exemple, que la **vitesse** est un type de différence spécifique à standardiser pour le champ sémantique des véhicules.

15. Cette hiérarchie sera développée à partir de celle déjà élaborée dans le cadre du projet de base de données lexicale DiCo, présentée dans (Polguère, 2003).

16. On peut envisager deux approches de la délimitation des **rôles**. Il est possible de considérer que ces rôles forment un ensemble restreint et prédéfini. C'est l'approche adoptée par le *Lexique Génératif* (Pustejovsky, 1995), qui propose une structure lexicale composée de quatre types de traits appelés *qualia* (`formal`, `constitutive`, `telic` et `agentive`). Ainsi, quelle que soit la partie du discours de l'unité définie et quel que soit son type sémantique, l'article représentant son sens sera constitué de ces quatre rôles, auxquels on attribuera ou non une valeur. On peut à l'inverse considérer, et c'est l'approche que nous adoptons, que ces rôles forment un ensemble fini, mais susceptible d'être si vaste qu'il ne pourra être connu qu'à l'issue de la description du lexique.

(14) REMORQUE (acception **B**)

```
<PARAPH>
  <CC étiq=véhicule>Bateau ou véhicule à roues</CC>
  <CP rôle=mode_fonctionnement>dépourvu d'un moyen de propulsion propre</CP>
  <CP rôle=fonction>et employé pour le transport des marchandises et/ou
      des voyageurs</CP>
</PARAPH>
```

(15) LANDAU (acception **B**)

```
<PARAPH>
  <CC étiq=véhicule>Voiture d'enfant</CC>
  <CP rôle=parties>munie de grandes roues, d'une capote pliante
       surmontant une caisse suspendue garnie d'une literie</CP>
  <CP rôle=fonction>et qui permet de promener un tout jeune enfant, notamment
      en position allongée</CP>
</PARAPH>
```

(16) BOLIDE (acception **B**)

```
<PARAPH>
  <CC étiq=véhicule>Véhicule (avion, automobile)</CC>
  <CP rôle=vitesse>allant à très grande vitesse</CP>
</PARAPH>
```

(17) TACOT (acception **B**)

```
<PARAPH>
  <CC étiq=voiture>Voiture</CC>
  <CP rôle=apparence>démodée</CP>
  <CP rôle=état>en mauvais état (mécanique, carrosserie)</CP>
  <CP rôle=son>qui fait du bruit</CP>
  <CP rôle=vitesse>et n'avance pas</CP>
</PARAPH>
```

3.5 Étape 3 : construction d'une base dérivée

La version entièrement structurée des définitions du TLFi pourra être utilisée comme « fonds lexicographique » à partir duquel sera entrepris un nouveau travail de rédaction d'articles pour le français contemporain, travail visant la production d'une base dérivée formelle exploitable informatiquement. La nomenclature de l'actuel TLFi contient un nombre important de lexies n'appartenant pas ou plus au français courant. Nous ne sélectionnerons, pour la nouvelle base de données, que les sens utilisés en français contemporain, en leur associant systématiquement des phrases d'exemple tirées de corpus de langue usuelle.

Une fois la nouvelle nomenclature mise en place, il nous faudra procéder à l'explicitation de la structure actancielle des lexies prédicatives, condition nécessaire pour une bonne définition de ces unités. Par exemple, nous indiquerons que le nom RANCUNE a trois actants : la personne X qui éprouve ce sentiment, la personne Y qui fait l'objet de ce sentiment et l'événement Z qui a suscité ce sentiment. Il est du reste important que les informations contenues dans une base de données sémantique puissent être mises en parallèle avec des informations morphologiques et syntaxiques. On pourra ainsi indiquer que l'adjectif RANCUNIER caractérise le premier actant du nom RANCUNE, ou encore que le second actant de ce nom peut être introduit par différentes prépositions (*rancune de X **vis-à-vis** de Y, rancune de X **à l'encontre** de Y*, etc.) [17].

L'explicitation de la structure actancielle des lexies servira de support à la réécriture des définitions dans un langage contrôlé. Certaines règles de réécriture pourront être appliquées de manière automatique. Par exemple, les composantes périphériques du type `fonction` (comme dans le cas des définitions de lexies étiquetées `véhicule`) seront systématiquement introduite par le syntagme *servant à* au lieu des différents *destiné à*, *qui sert à*, *spécialement équipé pour*, etc., relevés dans les définitions du TLFi. La plupart des règles de réécriture de définitions nécessiteront toutefois un travail important de la part des lexicographes, secondés par des applications d'aide à l'encodage des données (Barque & Nasr, à paraître). Notons, fi-

17. Voir les descriptions des liens de fonctions lexicales de la Lexicologie Explicative et Combinatoire.

nalement, que les formes lexicales utilisées dans les définitions seront désambiguïsées par le numéro de l'acception auxquelles elles correspondent dans la nouvelle nomenclature.

4 Conclusion

Le développement d'une base de données sémantique dérivée du TLFi est un projet ambitieux qui ne pourra se réaliser qu'au terme de plusieurs années de travail lexicographique. Les premières étapes de sa réalisation, qui seront menées à bien dans un délai beaucoup plus court, donneront déjà des résultats importants que nous énumérons ici.

- Valorisation de l'actuel TLFi : l'ajout d'information concernant la structure des définitions du TLFi rend la base encore plus attractive en permettant aux utilisateurs d'effectuer des requêtes plus fines que celles qu'il est possible d'effectuer actuellement. Par exemple, on pourra connaître l'ensemble des lexies du français dénotant un sentiment, ou encore connaître l'ensemble des lexies du français qui dénotent un véhicule et dont la fonction de ce véhicule est mentionnée dans leur définition, etc.
- Définition d'un format pour les définitions lexicographiques, dans le cadre du développement de lexiques pour le traitement automatique de la langue (TAL). En effet, la nouvelle norme ISO de structuration des données lexicales informatisées, *Lexical Markup Framework* (LMF), a défini un standard d'encodage des informations sémantiques (entre autres) des lexiques destinés au TAL, sans toutefois proposer de structuration **interne** pour les définitions lexicographiques (LMF, 2008) [18]. Ces dernières sont pourtant présentes dans les lexiques les plus utilisés en TAL — notamment WordNet (Fellbaum, 1998) et FrameNet (Fillmore et al., 2003), pour l'anglais — et mériteraient d'être rendues exploitables informatiquement suivant un modèle général standardisé de structuration.
- Une ressource pour la recherche en sémantique lexicale : la hiérarchie des étiquettes sémantiques développée pour le balisage des définitions du TLFi sera probablement complète avant même la fin du processus de structuration des définitions. Elle offrira une description des principaux sens lexicaux du français, organisés en hiérarchie, et indiquera les propriétés lexicograpiquement pertinentes associées à ces sens généraux.

Remerciement

Nous remercions Jean-Marie Pierrel et Pascale Bernard de l'ATILF pour leur aide sur le TLFi. Un grand merci aussi à Claudia Fecteau, Anne-Laure Jousse et Olivier Taïs pour leur implication dans le travail de balisage des définitions. Nous sommes très reconnaissants au réviseur de MTT'09 pour ses commentaires sur la première version de cet article. La recherche présentée ici est financée par des subventions du Fonds de recherche sur la société et la culture du Québec (FQRSC) et du Conseil de recherches en sciences humaines du Canada (CRSH).

Bibliographie

Altman, Joel, et Alain Polguère. 2003. La BDéf : base de définitions dérivée du *Dictionnaire explicatif et combinatoire*. *Proceedings of the First international conference on the Meaning- Text Theory (MTT'2003)*, Paris, 43–54.

Atkins, B. T. Sue and Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford University Press, Oxford, UK.

Barnbrook, Geoff. 2002. *Defining Language. A local grammar of definition sentences*. Benjamins, Amsterdam/Philadelphia.

Barque, Lucie. 2008. *Description et formalisation de la polysémie régulière du français*. Thèse de doctorat, Université Paris 7.

18. Notons qu'il est toutefois possible d'utiliser le cadre formel de LMF pour encoder divers types de structuration des définitions, comme celui de la BDéf (Francopoulo, 2005, pages 19–21).

Barque, Lucie, et Alexis Nasr. à paraître. Un modèle formel de descriptions lexicales : du formalisme BDéf aux structures de traits typées. *Traitement Automatique des Langues (T.A.L.)*, 50(1).

Barque, Lucie, et Alain Polguère. 2009. *Guide des annotateurs pour la structuration des dénitions du TLFi*. OLST, Université de Montréal.
[Accès en ligne : `http://www.olst.umontreal.ca/pdf/guideAnnoTLFi.pdf`]

Benson, Morton, Evelyn Benson, and Robert Ilson. 1986. *Lexicographic Description of English*. Language Companion Series 14, Benjamins, Amsterdam/Philadelphia.

Dendien, Jacques, et Jean-Marie Pierrel. 2003. Le Trésor de la Langue Française informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence. *Traitement Automatique des Langues (T.A.L.)*, 44(2) :11–37.

Dostie, Gaétane, Igor Mel'čuk, et Alain Polguère. 1999. Méthodologie d'élaboration des articles du dictionnaire explicatif et combinatoire du français contemporain. Dans I. Mel'čuk, N. Arbatchewsky-Jumarie, L. Iordanskaja, S. Mantha et A. Polguère (dir.) : *Dictionnaire explicatif et combinatoire du français contemporain : recherches lexico-sémantiques IV*, 11–28. Les Presses de l'Université de Montréal, Montréal.

Fodor, Jerry A., Merrill F. Garrett, Edward C. T. Walker, and Cornelia H. Parkes. 1980. Against definitions. *Cognition*, 8 :263–367.

Fellbaum, Christiane. 1998. *WordNet : An Electronic Lexical Database.* MIT Press, Cambridge, MA.

Fillmore, Charles J., Chris Johnson, and Miriam R. L. Petruck. 2003. Background to FrameNet *International Journal of Lexicography*, 16 :235–250

Francopoulo, Gil. 2005. *Extended examples of lexicons using LMF (auxiliary working paper for LMF)*. Rapport technique, INRIA-Loria.

Fontenelle, Thierry 2009. Linguistic research and learners dictionaries : the *Longman Dictionary of Contemporary English*. Dans A. P. Cowie (dir) : *The Oxford History of English Lexicography*, vol. II, 412–435. Oxford University Press, Oxford.

Frassi, Paolo. 2007. *La definizione nel* Trésor de la langue française *: studio tipologico e metalinguistico*. Thèse de doctorat, Università degli Studi di Verona, Vérone.

Kittredge, Richard I. 1982. Variation and homogeneity of sublanguages. Dans R. Kittredge and J. Lehrberger (dir) : *Sublanguage : Studies of Language in Restricted Semantic Domains*, 107–137. de Gruyter, Berlin.

Language resource management — Lexical markup framework (LMF). 2008. ISO/TC 37/SC 4 N453 (N330 Rev.16).
[Accès en ligne : `http://www.lexicalmarkupframwork.org`]

Lexis. Larousse de la langue française. 2002. Sous la direction de Jean Dubois. Larousse/VUEF, Paris.

Mel'čuk, Igor, et al. 1984, 1988, 1992, 1999. *Dictionnaire explicatif et combinatoire du français contemporain : recherches lexico-sémantiques I–IV*. Les Presses de l'Université de Montréal, Montréal.

Mel'čuk, Igor, André Clas, et Alain Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*. Duculot, Louvain-la-Neuve.

Nouveau Petit Robert. 2007. Sous la direction de Josette Rey-Debove et Alain Rey. Dictionnaires Le Robert, Paris.

Polguère, Alain. 1997. Meaning-Text Semantic Networks as a Formal Language. Dans L. Wanner (dir) : *Recent Trends in Meaning-Text Theory*, 1–24. Language Companion Series 39, Benjamins, Amsterdam/Philadelphia.

Polguère, Alain 2003. Étiquetage sémantique des lexies dans la base de données DiCo. *Traitement Automatique des Langues*, 44(2) :39–68.

Polguère, Alain. 2008. *Lexicologie et sémantique lexicale. Notions fondamentales*. Les Presses de l'Université de Montréal, Montréal.

Pustejovsky, James. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA & London, UK.

Riemer, Nick. 2006. Reductive Paraphrase and Meaning : A Critique of Wierzbickian Semantics. *Linguistics and Philosophy*, 29 : 347–379.

Rundell, Michael. 2008. More Than One Way to Skin a Cat : Why Full-Sentence Definitions Have Not Been Universally Adopted. Dans T. Fontenelle (dir) : *Practical Lexicography. A Reader*, 197–209. Oxford University Press, Oxford, UK.

Sinclair, John M. (dir.). 1990. *Collins Cobuild Student's Dictionary*. Collins, London & Glasgow.

Trésor de la Langue Française informatisé (TLFi). 2004. CNRS Éditions, Paris.
[Accès en ligne : `http://atilf.atilf.fr/tlf.htm`]

Wierzbicka, Anna. 1987. *English Speech Act Verbs. A Semantic Dictionary*. Academic Press, Sydney et al.

Wierzbicka, Anna. 1996. *Semantics : Primes and Universals*. Oxford University Press, Oxford, UK.

Domain, Domain Features of Lexical Functions, and Generation of Values by analogy according to the MTT approach

María Auxiliadora Barrios
Universidad Complutense de Madrid
auxiba@filol.ucm.es

Abstract

The purpose of the work here presented is to show the convenience of studying the Domain of Lexical Functions, that is, the paradigm of Lexical Units for which each Lexical Function was created. The definition of Domain of Lexical Function concept, coined but not developed in MTT, allows us to define the concept of Domain Features and to describe nouns attached to every LF. Both concepts provide a new working methodology, the Generation of Values by Analogy, and prove that many of the collocations are characterized by some kind of semantic motivations[2].

1 Introduction

In the same manner as the base of a collocation imposes some restrictions on the collocate (especially in support verbs), Lexical Functions (LFs) also impose restrictions on their arguments. These restrictions, which I refer to as domain features, can be defined when studying the domain of LFs, that is, the paradigm of the lexical units for which each LF was defined. I have described the domains and defined the domain features of those LFs which are the most productive in Spanish: fulfillment verbs, support verbs, phasal verbs and causative verbs.

With the aim of carrying out this study, it was necessary to collect a considerable amount of data in Spanish. In order to obtain examples efficiently, the principle of 'lexical inheritance' (Mel'čuk & Wanner, 1996) was extended and applied to the domain of LFs. In the research work here presented, the *generation of values by analogy* is proposed based both on the semantic motivation of many collocations and the systematization of those lacking motivation. The extension of this principle maintains that those lexical units sharing a semantic label can inherit their LFs values[3] automatically, as (1)-(3) shows:

(1) **Real$_1$**('clothes') = **Real$_1$**(skirt) = **Real$_1$**(trouser) = **Real$_1$**(shirt) = *to wear*
(2) **IncepReal$_1$**('clothes') = **IncepReal$_1$**(skirt) = **IncepReal$_1$**(trouser) = **IncepReal$_1$**(shirt) = *to put on*
(3) **FinReal$_1$**('clothes') = **FinReal$_1$**(skirt) = **FinReal$_1$**(trouser) = **FinReal$_1$**(shirt) = *to take off*

In order to implement and validate this proposal, a database has been designed. This database, known as *BADELE.3000* (Barrios & Bernardos, 2007), contains 3,300 nouns, which are the most frequent in the Spanish language spoken in Spain. With this database, approximately 9,000 lexical relations were obtained automatically. Moreover, it was possible to add some other lexical relations found in the combinatorial dictionaries of Spanish (Bosque, 2006; 2004). In fact, a total of 20,700 relations were formalized by means of LFs.

The directionality of the lexical selection seems to be conflictive. The MTT approach defends that the base of a collocation selects the predicate, while Bosque's approach (2004) proposes that it is the

[2] This work has been partially supported by the National Project "GeoBuddies: Anotación semántica colaborativa con dispositivos móviles en el Camino de Santiago" (TSI2007-65677C02).
[3] For the concepts glosses and default values see Alonso Ramos (2006).

predicate that selects its arguments. The study of the domain of LFs demonstrates that these opinions are complementary. In fact, *lexical selection* is used with a different sense in each approach. For Meaning Text Theory, its sense is affected by the predominance of a communicative point of view. For Bosque, the grammatical point of view is the predominant one. Precisely, the study of the glosses and the values of the LFs here developed integrates both approaches: glosses and values are selected by their arguments, but at the same time, when the meaning is preserved, they impose semantic restrictions on their arguments.

The structure of the paper is organized as follows. Section 1 defines the domain of LFs and the generation by analogy of the values of LFs. Section 2 describes fulfillment domain features. Section 3 describes support domain features. Section 4 describes the phasal domain features. Section 5 describes some causative domain features. Section 6 deals with the problems of semantic motivation in collocations and the directionality of lexical selection**.** Finally, section 7 summarizes the conclusions.

2 Domain of LFs, Domain features and generation by analogy of the values of LFs

Lexical Functions were introduced and developed within the Meaning – Text theory (Mel'čuk, 1996). Among the mathematical concepts related to *function*, it is interesting to pay attention to the term *domain*. In the mathematical field, it is the set of all the possible values that a function is defined for. Similarly, as Mel'čuk proposes,

> The *domain of the lexical functions* is here understood as the set of LUs for which each LF was defined, in other words the set of arguments of every LF.
>
> **Definition 1: Mel'čuk's (1996:76) definition of LF Domain**

For instance, **Oper$_1$** domain contains lexical units (LUs) such as *lecture* (*to give a lecture*), *class* (*to teach/ hold a class*), *walk* (*to have a walk*), *support* (*to lend support*), *control* (*to have control over*), etc. LF domain must be understood as a potential domain: the set of LUs for which the LF has a potential meaning, but not necessarily a value.

As the next sections will show, most of LFs impose semantic conditions on their keywords, so LFs domain features can be defined as follows.

> *LFs domain features* are semantic features essential for each LF keyword, in so far as a LF calls this features for their keywords. One or more features can be characteristic of every LF.
>
> **Definition 2: LFs Domain features**

Many LFs domains contain one or more semantic field. Other approaches, different from the MTT approach, such as Pustejovsky's generative lexicon (1995) and WordNet's lexical database[4] are based on *semantic inheritance*, i.e. the semantic classification of lexemes (for example, *cat*, *dog* and *horse*) into a class (in this case, ANIMAL) which includes all the semantic features they share. Mel'čuk and Wanner (1996) apply the semantic inheritance to the lexicographic representation and propose the *Lexical Inheritance Principle*[5], which they implemented and validated for the lexical field of emotion in German. Thus, it will be possible to foretell that the verb *to feel* combines with every lexical unit meaning 'emotion'. Sanromán (2003) employed the lexical inheritance principle[6] applied in the *DiCE* (Alonso

[4] http://wordnet.princeton.edu/

[5] Although the principle of lexical inheritance had already been applied to some entries of the *DECFC* (cf. , *DECFC* vol. IV, p. 166, the entry *carotte$_{1c}$* relating to *legume*), Mel'čuk and Wanner (1996) were the ones who introduced it. Their idea could be summarized by the paragraph: "The semantic features of words that share some values for certain syntagmatic LFs can be analyzed. It must be possible to generalize restricted lexical co-occurrence instantiations along semantic lines".

[6] Chapter 3 contains a description of Spanish emotion nouns based on the concept of lexical inheritance.

Ramos and Sanromán: 2000), which is a database based on the MTT principles[7]. Both authors followed Mel'čuk and Wanner's methodology and applied it to the same lexical field, viz. emotions, in Spanish. The *DiCE* contains some shared values for some LFs. As in German, the verb *sentir* (to feel) is a shared value for **Oper$_1$** for emotion Spanish words. Sanromán (2003:89) proved that collocations are at least partially semantically motivated. Outside the framework of the MTT, such a principle has been implemented in a certain way by Bosque (2006) in the *Diccionario Combinatorio Práctico del Español* (Practical Combinatory Dictionary of Spanish), which contains common collocatives for many different entries grouped in 18 generic ones.

It is possible to apply the Lexical Inheritance Principle in more than one way because there are several possible generalizations about values (Mel'čuk, 1996: 76-78), among others, the values of the LFs can be shared by the keywords (as *to feel* by emotion nouns, or *to drive* by motor-vehicle nouns). But, in my view, it is also possible to generalize about LFs themselves, because they impose restrictions on their keywords, and, consequently, the keywords share some semantic properties. Actually, the keywords belonging to the same LF domain usually share both domain features and the semantic label (the semantic label corresponds to the first part of the definition of the *keyword*, as 'animal' and 'food' for *lamb*, see Polguère, 2003:43).

In this sense, there is another way of application of the Lexical Inheritance Principle not mentioned by Mel'čuk: the notion of the domain of a LF opens a way to propose an extension of the *Lexical Inheritance Principle*, what I call the *generation by analogy of the values of LFs*:

> Values of LFs often correlate with semantic features of their keywords. Therefore, LUs belonging to the same LF domain and having the same semantic label often share LF values.
>
> **Definition 3: Generation by analogy of the values of LFs**

The *generation by analogy of the values of LFs* is the key-concept for this work. The use of this principle allows us to get several Spanish values of Lexical Functions for some lexical fields automatically, as (1)-(3) showed. However, there is another way to implement the generation, by paraphrases of some LFs. For example, the LF called **CausFunc$_0$** means 'to cause something to begin to exist'. Then the semantic labels of the nouns that can be combined with a verb that means 'to cause something to begin to exist' were selected, among others, *ropa* (clothes), *obra artística* (artistic work), *vivienda* (accommodation), *energía* (energy). They were related to the lexical units that express **CausFunc$_0$** sense –for instance, *confeccionar* (to make clothes), *componer* (to compose artistic works), *construir* (to build accommodation), *producir* (to produce energy), etc. All the lexical units under these semantic labels inherited automatically these verbs, as (4)-(9) shows:

(4) **CausFunc$_0$** = 'causar que algo empiece a existir'
CausFunc$_0$ = 'to cause something to begin to exist'

(5) 'Causar que algo empiece a existir' = confeccionar, componer, construir, producir, etc.
'To cause something to begin to exist' = to make, to compose, to build, to produce, etc.

(6) **CausFunc$_0$**(ropa) = *confeccionar* > confeccionar una camiseta/ pantalones/ falda, etc.
CausFunc$_0$(clothes) = *to make* > to make a T-shirt/ trousers/ skirt/ etc.

(7) **CausFunc$_0$**(obra artística) = *componer* > componer un poema/ libro/ argumento, etc.
CausFunc$_0$(artistic work) = *to compose* > to compose a poem/ book/ plot/ etc.

(8) **CausFunc$_0$**(vivienda) = *construir* > construir una casa/ rascacielos/ apartamento, etc.
CausFunc$_0$(accommodation) = *to build* > to build a house/ skyscraper/ apartment/ etc.

(9) **CausFunc$_0$**(energía) = *producir* > producir luz/ gas/ petróleo, etc.
CausFunc$_0$(energy) = *to produce* > to produce light/ gas, petrol, etc.

[7] Diccionario de colocaciones del español. Alonso Ramos. http://www.dicesp.com

Finally, other verbs from the Spanish combinatory dictionaries (Bosque: 2004; 2006) were added (*escribir, esbozar, crear un poema*; *to write/ draft/ create a poem*, etc.). The implementation of the domain inheritance principle allowed us to obtain automatically more than 9,000 collocations and, at the same time, to validate the principle itself.

This principle allows to formulate very plausible hypotheses, which are later checked.

On the following pages, some collocations which were obtained automatically will be shown, as well as the domain features defined for each family of LFs.

3 Fulfillment verbs

The values of LFs **Real$_{0/I}$**, **Fact$_{0/I}$** and **Labreal$_{ij}$** are *fulfillment verbs* and mean 'to fulfill the requirement of L', or 'to do with L what you are supposed to do with L', or 'L fulfils its requirement' (Mel'čuk, 1996, 68). Mel'čuk's examples of the collocations covered by these LFs are *to prove an accusation, to drive a bus*, *to ride on bus, to take hint, to come true (hope)*. The implementation of the generation by analogy (we found 577 relations between semantic labels and fulfillment LFs, such as *to put the clothes on, to take the clothes off, to wear the clothes*, etc.) allowed us to arrive automatically at a paradigm of 3,995 Spanish examples. After adding those collocations that were obtained non automatically, there was a total of 5,171. Some examples are: *el coche no funciona, the car doesn't work; la ley surtió los efectos deseables, the law produced effects we find desirable; segar (algo) con la hoz*, *to reap (something) with the sickle*; *barrer (el suelo) con la escoba*, *to sweep (something) with the broom*; *escribir (algo) en un folio*, *to write (something) on a sheet of paper.*

ULs belonging to the fullfilment domain have some common semantic characteristics which allow defining fullfilment domain features: 'finality', 'utility' and 'expectability'. The two first features are present on most of the LUs, such as *the bullet hit her on the arm* (the finality of a bullet is to hit or kill), *a flute sounds* (the finality of a flute is to sound) and *an apple tree that doesn't bear fruit* (the utility of a fruit tree is to bear fruit). The last feature is typical of abstract nouns, such as *admiración* (admiration)*, memoria* (memory)*, recuerdo* (memory, remembrance)*, juicio* (trial), for which this feature is expressed by the collocations *mostrar admiración* (to show admiration)*, retener en la memoria* (to keep it in one's head)*, guardar los recuerdos* (to keep it in one's head)*, ganar un juicio* (to win a court case).

4 Support verbs

The values of LFs **Oper$_i$**, **Func$_{0/i}$** and **Labor$_{ij}$** are called *support verbs*. They are not necessarily semantically full, but frequently emptied. Their keyword presupposes actants, they are predicates (Mel'čuk, 1996, 68). Some of Mel'čuk's examples of the collocations covered by these LFs are *to deal a blow, to receive a blow from, blow comes from, blow falls upon, to put up resistance, to meet resistance, to give an order, to receive an order.*

In contrast with fulfillment LFs, because of support verbs' lack of meaning, it is not possible to arrive at some support verbs paraphrases and to foretell the values of these LFs. However, most of the support verb collocations are equivalent to some other verb[8], as it is the case of *to bang/ beat/ hit* and *to deal a blow*, *to resist* and *to put up resistance*, *to order* and *to give an order*. Then it is at least possible to predict the keywords of these LFs: nouns whose meaning is equivalent to the meaning of some verbs (*blow/ to beat*; *resistance/ to resist*; *order/ to order*). All these nouns were extracted from the database (they had been grouped previously by semantic label). The study of these Lexical Units allowed to predict the support verb domain[9] and to define the support verbs domain features[10]. Then 1331 support verbs collocations

[8] Our database contains 777 support verb collocations; more than 700 have an equivalent verb.

[9] Due to lack of space, it is not possible to describe how it was predicted, but first, we listed **Oper$_i$** keywords (by the existence of equivalents verbs), and after we checked which of them were **Func$_i$** and **Labor$_{ij}$** keywords. In order to know the differences between each support verb LF, see Mel'čuk, 1996.

[10] It is not possible to predict LUs belonging to the **Labor** domain. I found few examples in our database: *someter a alguien a juicio, someter a juicio, llevar a alguien a juicio, dejar algo/alguien en el abandono, tener algo en venta.*

were included in our database, 993 of them obtained automatically by 134 pairs of semantic labels and default values.

Func$_0$ keywords were excluded in the process. I think that **Func$_0$** is a LF different from **Func$_i$**, **Oper$_i$** and **Labor$_{ij}$**, because there is a paraphrase of the **Func$_0$** meaning ('it does exist'), but it is not possible to find any meaning for the other three LFs, but default glosses (Alonso Ramos, 2006). Differences among **Func$_i$**, **Oper$_i$** and **Labor$_{ij}$** are purely syntactic, as shown by the fact that most of the **Func$_i$** and **Labor$_{ij}$** keywords are also **Oper$_i$** keywords, while many of **Func$_0$** keywords are not always **Oper$_i$** keywords[11].

To my knowledge, the domain feature of **Func$_0$** is 'something that exists doing something in a specific way', or more simply, 'general situation', feature which is present in keywords such as *lluvia (la lluvia cae*), *rain* (*the rain falls*); *viento (el viento sopla)*, *wind* (*the wind flows*); *fuego (el fuego quema), fire* (*the fire burns*); *luz (la luz brilla)*, *light (the light shines*). However, domain features of **Func$_i$**, **Oper$_i$** and **Labor** are: a) 'action that someone/something realizes' (present in words such as *paseo* (*walk*), *advertencia* (*warn*), *ocupación* (*occupation*), *orden* (*order*), *grito* (*cry*), which produces collocations such as *dar un paseo, to go for a walk,* etc.); b) 'someone/ something's property' (present in words such as *capacidad* (*ability*), *inteligencia* (*intelligence*), *simpatía* (*friendliness*), *amplitud* (*spaciousness*), which produces collocations such as *tener capacidad*, *to have an ability*, etc.); c) 'someone/ something's situation' (present in words such as *frío* (*cold*), *desgracia* (*misfortune*), *choque* (*collision*), which produces collocations such as *tener frío*, *to be cold*, etc.)

5 Phasal verbs

The values of LFs **Incep**, **Cont** and **Fin** are called *phasal verbs*. They are semantically full, and mean 'to begin', 'to continue' and 'to finish', respectively. Their keywords are verbs or verbal LFs, such as fulfillment and support verbs LFs (more details in Mel'čuk, 1996, 64-65). Some of Mel'čuk's examples of the collocations covered by these LFs are *to open fire* and *to fall under the power of,* for **Incep**; *to retain one's power over*, *the offer stands* for **Cont**; and *to lose one's power* for **Fin**.

Because of their semantic properties, phasal LFs impose on their keywords only the grammatical condition 'to be a verb'. This explains why these LFs combine with fulfillment and support LFs. This peculiarity implies that phasal verbs domains are a subdomain of the fulfillment and support verbs domains. For instance, *joy* and all the nouns of feelings belong to the phasal domain (*nace un sentimiento*, *a feeling is born*) and, as it was mentioned above, also to the support verb domain (*sentir alegría*, *to feel joy*)[12]. Consequently, the semantic labels of fulfillment and support verb domains were reviewed in order to select those for which it is possible to predicate 'it begins P', 'it continues P', 'it finishes P', where P is a fulfillment or a support verb. Then, the most productive phasal LFs (**IncepFunc**, **FinFunc** and **IncepReal**) were inherited. However, in my view **IncepFunc$_0$**, **ContFunc$_0$**and **FinFunc$_0$** are different from the rest of phasal LFs, because their domains are independent of fulfillment or support verbs, and it was possible to foretell their keywords by LFs meanings ('to start/ continue/ finish to exist'), and to arrive at *erosion starts/ finishes, hate comes out/ melts away, love is born/ disappears, light arises/ dies out, the sunset arrives*. More than 3,600 phasal verb collocations were obtained, 1,277 of them automatically by 154 pairs of semantic labels and glosses.

Domain features of phasal LFs were described; they allow us to distinguish and define five kinds of phasal nouns (all of them are for situation nouns): *period, state, punctual nouns, lasting nouns, process nouns*.

---

[11] In our database there are some examples of **Func$_0$** collocations which have keywords that are also **Oper$_1$** keywords, such as *consumarse una desgracia* (*a misfortune happens*): its **Oper$_1$** collocation is *caer en/ sufrir una desgracia* (*to fall from favour*); *collision*, *producirse un choque* (*a collision is on*), its **Oper$_1$** collocation is *sufrir un choque* (*to suffer a collision*); *tener lugar una revolución* (*a revolution takes place*), its **Oper$_1$** collocation is *hacer una revolución* (*to make a revolution*). Some **Func$_0$** keywords that do not belong to **Oper$_1$** domain are *fuego* (*fire*); *luz* (*light*); *viento* (*wind*); *lluvia* (*rain*); *ola* (*wave*); *alba* (*daybreak*); *trueno* (*thunderclap*); *rayo* (*ray*), etc.

[12] Fontenelle (98) describes words like *anger* and *enthusiasm* in terms of LFs showing that they are phasal verbs with a beginning, a continuation phase, a culmination and an end.

1) *Punctual nouns* design instantaneous phenomena such as *rayo* (*ray*), *chispa* (*spark*), *chirrido* (*creaking*). They do not combine with *durar* (*to last*) and *durante* (*during/for*), and for these LUs, LFs **IncepFunc$_0$** and **Func$_0$** are expressed by the same values (this phenomenon is called overlapping of LFs), as it occurs in *caer un rayo* (*the lightning struck*), *saltar una chispa* (*a spark flew*). 2) *Lasting nouns* design events that can be a bit lasting, such as *tormenta* (*storm*), *enfermedad* (*illness*), *crisis* (*crisis*). They combine with *durar* (*to last*) and *durante* (*during/for*). 3) *Periods* are nouns of limited duration, such as *minuto* (*minute*), *mes* (*month*), *año* (*year*). They only combine with *durante* (*during/for*). 4) *States* are nouns that design events that can be quite lasting (their meaning doesn't demand necessarily an end), such as *paz* (*peace*), *matrimonio* (*marriage*), *fama* (*fame*); they only combine with *durar* (*to last*). 5) *Process nouns* design events that can be a bit lasting, such as *alba* (*daybreak*), *anochecer* (*night fall*), *oxidación* (*oxidation*). Like lasting nouns, they combine with *durar* (*to last*) and *durante* (*during/for*); and like punctual nouns, there is an overlapping of LFs **IncepFunc$_0$** and **Func$_0$**, *llegar el anochecer* (*nightfall arrives*), *producirse la oxidación* (*oxidation is on*).

Domain features of phasal verbs for concrete nouns were not found, although they could be expressed approximately as 'that can be begin/ continue/ finish P', where P is a fulfillment verb. However, this feature does not depend necessarily on linguistic characteristics: for example, the expression *to put on a hat* exists because in our world, it is the beginning action for wearing a hat.

6 Causative verbs

The values of LFs **Caus**, **Perm** and **Liqu** are called *causative verbs*. They are semantically full and mean 'to cause', 'to allow' and 'to liquidate', respectively. Their keywords are verbs or verbal LFs, such as fulfillment and support verbs LFs (more details in Mel'čuk, 1996, 65-68). Some of Mel'čuk's examples of the collocations covered by these LFs are *to lead opinion, to raise hope in,* for **Caus**; *to condone an aggression* for **Perm**; and *to stop an aggression, to wipe out the traces* for **Liqu**.

In the same way that phasal LFs, causative LFs impose on their keyword the grammatical condition 'to be a verb'. Consequently, causative verb domains are also a subdomain of fulfillment and support verb domains. However, in my view, **CausFunc$_0$**, **PermFunc$_0$** and **LiquFunc$_0$** are different from the rest of causative LFs. They contain phasal LFs inside, as show their meanings, 'to cause that something starts to exist', 'to cause that something continues to exist' and 'to cause that something finishes to exist', respectively. That explains why, as in the case of phasal LFs, it was possible to foretell their domains.

In the same manner as in section 4, semantic labels of fulfillment and support verb domains were reviewed in order to arrive at the rest of causative LFs keywords. In my view, **CausFact$_0$** can be paraphrased as 'to cause + verb', where the verb is the value of **Fact$_0$**: for instance, *to play a guitar* means 'to cause the guitar to sound' (*to sound* is the **Fact$_0$** value). To my knowledge, it is possible to find similar paraphrases for causative fulfillment LFs.

It was not possible to define causative domain features, except for two groups of LFs. The first one is **CausFunc$_0$**, **PermFunc$_0$** and **LiquFunc$_0$**, the keywords of which are predictable. Their feature is 'there is an external or an internal agent that can cause that this exist". We have proposed the name of *causative noun* only for those nouns belonging to these domains. These are nouns that combine with verbs having a first argument that is the 'causer' of what is designated by the noun.

There is a second one group, the group of LF **Caus$_2$Func$_1$**, that is quite productive on emotions lexical field, such as *despertar admiración* (*to arise admiration*), *inculcar cariño* (*instill affection*), *provocar celos* (*to provoke jealousy*). Its domain feature is 'Y can provoke that X feels this emotion', where X and Y are the first and second actant of the keyword.

7 Semantic motivation and directionality of lexical selection

The semantic motivation of LFs is not well accepted by most of MTT scholars, because there are abundant examples that show how synonym bases select different values, such as *dar un beso* (to give a kiss) /**dar una caricia* (to caress); *hacer una advertencia* (to give a warning to someone)/ **hacer un*

aviso (to give someone an advice); *tomar una resolución/*tomar el propósito* (to adopt a resolution); *hacer un préstamo* (to give a loan to somebody)*/ *hacer una ayuda* (to help) (see Alonso Ramos, 1998).

In spite of that, according to Apresjan, semantic motivation of LFs is present even in support verbs LFs: "motivation is sufficiently great to provide a foundation for stating the general trends and forming useful lexicographic expectations, although it is insufficient for formulating rules" (Apresjan, Glovinskaja, 2007). Empirical evidence corroborates Apresjan's hypothesis, as his examples shows, such as *to carry on* + 'activities' (*research, work*), *to undergo* + 'processes' (*a change, decomposition, deformation*), *to be in* + 'state' (*confusion, despair, ecstasy*), etc. As Apresjan says, theoretical evidence comes from "the fact that the values of any collocate LF (…) obey the general laws of semantic agreement between collocating items in the text. But semantic agreement consists in the recurrence of a certain semantic component in the meaning of two collocating items".

Apresjan's evidence in favour of semantic motivation is in accordance with Bosque (2006, 2004) works. Bosque's two dictionaries follow the sense collocative > base, whereas MTT dictionaries follow the sense base > collocative. Bosque thinks the collocatives keep their meanings, and he points out that the collocatives impose restrictions on their bases. That explains why on his dictionaries nouns entries were obtained automatically, while verb, adjective and adverbial entries were written manually.

Bosque argues that the direction base > collocative is justified by what he calls *categorización multiple* (multiple categorization): verbs, adjectives and adverbials shows different meanings. For instance, the verb *leer* (to read) is a speech verb, as proved by the combination *leer en voz alta* (to read something aloud); this collocative is shared by other Spanish speech verbs, such as *hablar* (to speak)*, decir* (to say)*, quejarse* (to complain). But *leer* is also a perception verb, as proved by *leer de refilón* (out of the corner of one's eye); this expression can be combined with other Spanish perception verbs, such as *ver* (to see)*, mirar* (to look)*, tocar* (to touch). And finally, *leer* is also a consumption verb, as proved by *leer compulsivamente* (to read compulsively); this adverb combines with other Spanish consumption verbs, such as *comer* (to eat)*, fumar* (to smoke)*, comprar* (to buy).

Sections 3 and 4 of this work indicate that domain features of fulfillment and support verbs are semantic features: the meanings 'finality', 'utility' and 'expectability' correspond to the fulfilment verbs and the meanings 'someone/something's action', 'someone/ something's property', and 'someone/ something's situation' correspond to support verbs. As it was shown in section 5, it is not possible to define phasal domain features by meanings, but by grammatical properties[13]. Likewise, in section 6 only two causative domain features were defined. However, in my view, the fact that fulfillment and support verbs domain features are semantic features holds up the hypotheses of the semantic motivation of many of the collocations (as it was explained, phasal and causatives domain are known thanks to the fulfillment and support verbs domain).

Empirical evidence of semantic motivation was also obtained on the database *BADELE.3000* applying the domain inheritance principle. Apart from the examples already mentioned (where families of nouns semantically close share values of LFs), there are some other data that can corroborate the same hypotheses. For instance, *caer* (to fall down?) selects as first argument a noun with a meaning that requires the sense of 'moving down', such as *telón* (curtain)*, bomba* (bomb)*, hoja* (leaf)*, lluvia* (rain). On the other hand, these nouns need a verb in order to express different LFs meanings, such as 'to fulfil the requirement of' (which is the meaning of $\mathbf{Fact_0}$), *caer el telón* (the curtain comes down); 'to begin fulfilling the requirement of' (which is the meaning of $\mathbf{IncepFact_0}$), *caer las bombas* (the bombs fall); 'empezar a degradarse' (which is the meaning of **Degrad**), *caer las hojas* (the leafs fall); 'existir' (which is the meaning of $\mathbf{Func_0}$), *caer la lluvia* (the rain comes down). In all these collocations *caer* preserves the meaning 'to fall'. However, there are other nouns that select *caer* in order to express some particular LF meaning without the meaning 'to fall'. That is the case of *caer enfermo* (to become ill), 'to begin being

[13] For phasal nouns it is possible to predicate 'it begins P', 'it continues P', 'it finishes P', where P is a fulfilment or support verb. About causative domain features, see that only the meanings 'Y can provoke that X feels this emotion' ($\mathbf{Caus_2Func_1}$ feature), and 'there is an external or internal agent that can cause that this exist', ($\mathbf{CausFunc_0}$ feature) were defined.

ill' (which is the meaning of **IncepOper$_1$**), and *caer un imperio* (the empire crumbles), 'to finish existing' (which is the meaning of **FinFunc$_0$**).

This example shows how collocates can *select* and *be selected* at the same time: when collocates preserve their meaning, they select their arguments ('fall' selects what falls), and, at the same time, these arguments need a verb expressing some LF meaning and select this verb ('curtain' selects *fall*); when collocates do not preserve their meaning, they are selected by the base. Lack of space makes it difficult to mention other examples, but many of the verb entries in Bosque (2004, 2006) have similar properties.

As Apresjan claims, there are combinatory tendencies even in support verbs collocations. For instance, movement nouns usually combine with the verb *dar* (to give), as show *dar un paseo* (to walk)/ *una vuelta* (to go around something)/ *un giro* (turn)/ *un paso* (to take a step)/ *un brinco* (to jump)/ *un respingo* (to give a start)/ *un salto* (to jump)/ *una zancada* (stride), and with movement verbs related to driving a car, such as *dar un volantazo* (to swerve)/ *frenazo* (to slam on the brakes)/ *acelerón* (to put one's foot down).

8 Conclusions

The *domain of the LF* is a concept of great interest since it makes it possible to work within the framework of MTT and, at the same time, to establish a link with Bosque's approach. I have called *domain features* the semantic features shared by the keywords selected by each LF. The study of the domain of LFs and their domain features is based on the semantic motivation of those verbs which preserve their meaning, such as *caer* (fall), *cumplir* (fulfil), *obedecer* (obey); on the combinatory tendencies of semantically empty verbs such as *dar paseos* (go for a walk)/ *vueltas*, *giros* (turns)/ *saltos* (jumps); and also on verbs equivalent to keywords of support verbs.

There are values associated to the various LFs, so it was highly convenient to generate the inheritance of such values automatically, taking into account the LFs and using semantic labels. Thus, an extension of the *principle of lexical inheritance* (Mel'čuk and Wanner, 1996) was applied to the domains of LFs, and then the *generation of values by analogy* was defined. This extension of the principle allows two kinds of predictions: a) a given noun belongs to the domain of a LF; b) it has a certain value associated.

Fulfillment verb LFs need a thorough revision, because they express meanings that are not only related to lexis but also to semantics and pragmatics. Fulfillment domains verify that there is an objective feature that can be "expected" in a substantial amount of the collocations covered by these LFs. The problem lies on the fact that such feature varies in certain linguistic contexts due to extralinguistic reasons. Domain features of these nouns (Ls) were defined: 'finality' and 'utility' for concrete nouns, and 'expectability' for abstract nouns.

One key finding in support verbs is that among these LFs, **Func$_0$** differs semantically from **Oper** and **Labor**. The domain feature imposed by **Func$_0$** is 'L exists by doing something in some way', as in *cae la lluvia* (rain falls). However, **Oper** and **Labor** require their bases to be predicates usually having an equivalent verb. The feature domain that **Oper, Func$_i$** and **Labor** impose is an 'action performed by somebody', a 'property that somebody has' or a 'situation that is experimented'.

Some situation nouns from the support verb domain (*esbozar una sonrisa*, to crack a smile; *lanzar una carcajada*, burst into laughter) and some situation nouns form the fulfillment verb domain (*arrancar un motor*, to start the engine; *prender la cerilla*, to light a match, *pararse el tren*, stop (the train)) belong to the phasal verbs domain. The domain features of phasal nouns of support verbs domain are the following: 'it can last' (nouns with this feature combine with *durar* (to last), 'it takes place during/for' (nouns with this feature combine with *durante* (*during/for*) and 'it is punctual' (for which there is an overlap between **IncepFunc$_0$** and **Func$_0$**). The presence or absence of each of these domain features was useful to distinguish *punctual nouns*, *lasting nouns*, *periods*, *states* and *processes*.

Situation and entity nouns also belong to the causative LFs domain, which is too wide to give a precise definition of its domain features. For **CausFunc$_0$**, **PermFunc$_0$** and **LiquFunc$_0$**, the feature 'there is an internal or external agent that can make that exist' was proposed, and for **Caus$_2$Func$_1$**, the feature 'Y can make this affect X'.

The domain features of fulfillment and support verbs are semantic features, thus integrating with the Apresjan and Bosque theoretical evidences in favour of semantic motivation. The results of *BADELE.3000* show that it is possible to apply the *domain inheritance principle*, which is also based on semantic motivation. These data prove that there are some tendencies even in support verbs, and that collocatives can both select and be selected at the same time, that means that *selection* can be understood as a communicative process (MTT sense) or as a grammatical characteristic (Bosque' sense).

References

Alonso Ramos, Margarita. 2006. "Glosas para las colocaciones en el Diccionario de colocaciones del español. Diccionarios y Lexicografía". Anexos de Revista de Lexicografía 3. Universidade da Coruña: A Coruña, 59-88.

Apresjan, Jury; Glovinskaja, Marina J. 2007. Two Projects: English ECD and Russian Production Dictionary. In Gerdes *et al* (ed.), 53-68.

Barrios Rodríguez, Mª Auxiliadora; Bernardos, Mª Socorro. 2007. "*BaDELE.3000*: An implementation of the lexical inheritance principle". In Gerdes *et al*, 97-106.

Bosque, Ignacio. (dir.) 2006. *Diccionario práctico combinatorio del español contemporáneo*, Madrid: ed. SM.

Bosque, Ignacio. (dir.) 2004. Redes. Diccionario combinatorio del español contemporáneo, Madrid, ed. SM.

Cowie, Anthony P. (ed.). 1998. *Phraseology: Theory, Analysis and applications*. Oxford. Clarendon Press.

Fontenelle, Thierry. 1998. "Discovering significant lexical functions in dictionary entries". In Cowie (ed.), 189-207.

Gerdes, Kim; Reuther, Tilmann; Wanner, Leo (eds.). 2007. *Meaning-Text Theory 2007. Proceedings of the 3nd International Conference on Meaning-Text Theory. Wiener Slawistischer Almanach*. Sonderband, 69.

Mel'čuk, I., Wanner, L. 1996. "Lexical functions and lexical inheritance for emotion lexemes in German". In Wanner, L. (ed.), 209-278.

Mel'čuk, Igor. 1996. "Lexical functions: A tool for the description of lexical relations in a lexicon". In Wanner, L. (ed.), 37-102.

Polguère, Alain. 2003. "Étiquetage sémantique des lexies dans la base de données DiCo". *TAL*. Vol 44, nº 2/2003, 39-68.

Pustejovsky, James. 1995. *The generative lexicon*. Cambridge (Massachussets), London. MIT Press.

Sanromán Vilas, Begoña. 2003. *Semántica, sintaxis y combinatoria léxica de los nombres de emoción en español.* Tesis doctoral. Universidad de Helsinki.

Wanner, Leo (ed.) 1996. *Lexical functions in lexicography and natural language processing*. Amsterdam/ Philadelphia. John Benjamin.

Thematicity in Lushootseed syntax

David Beck
University of Alberta
4-45 Assiniboia Hall
Edmonton, AB T6E 2G7, Canada
dbeck@ualberta.ca

Abstract

The surface form of sentences in Lushootseed is governed primarily by considerations of Communicative Structure, in particular Thematicity. This paper will argue that Thematicity, rather than part of speech, governs the selection of the syntactic predicate in Lushootseed, and will offer some preliminary formal treatments of the phenomenon in terms of ranked constraints governing the transition between SemR and DSyntR. It will also be shown that the typological distinction between Lushootseed and more familiar languages like English can be accounted for by variable rankings of a shared set of constraints on this transition.

The surface form of sentences in the Salishan language Lushootseed, as in other languages in its family, is conditioned to a remarkable degree by Communicative Structure (Mel'čuk, 2001, a.k.a. Information Structure—Lambrecht, 1994; Vallduví, 1992). While previous attempts to come to terms with this aspect of Salishan syntax have made use of concepts such as Topic and Comment (Davis and Saunders, 1978) and Discourse Topic (Beck, 2000; Kinkade, 1990), this paper will attempt to account for a wider range of the effects of Communicative Structure on Lushootseed syntax by applying the model of Semantic Communicative Structure (Sem-CommS) outlined by Mel'čuk (2001). It will be argued that, contrary to traditional approaches to syntax, which give *a priori* primacy to lexical and syntactic categories in clause structure, Lushootseed requires that precedence be given to Communicative Structure in the organization of the clause. Some preliminary steps will also be taken towards formalizing a system of ranked constraints on the expression of particular elements of the Sem-CommS in Deep Syntactic Structure, and it will be shown that at least some of the typological differences separating Salishan from more familiar languages can be accounted for by the differential ordering of the same basic set of constraints.

1 Lushootseed clause structure

Lushootseed, like other Salishan languages, uses a system of pronominal clitics, agreement-markers, and a fairly rigid VSO word-order to encode grammatical relations (Czaykowska-Higgins and Kinkade, 1998). The simple transitive clause is illustrated in (1):[1]

(1) a. ʔu–gʷəč'–t čəd ti sqʷəbayʔ
PFV–seek–ICS 1SG.SUB DEF dog
'I sought the dog'

[1] The abbreviations used in glosses are: 1,2,3 = first-, second-, third-person; ADD = additive; ATTN = attenuative; DEF = definite; DC = diminished control; DIST = distal; ICS = internal causative; INTJ = interjection; NDEM = non-demonstrative; PASS = passive; PFV = perfective; PL = plural; PR = preposition; PROX = proximal; REM = remote; REFL = reflexive; SG = singular; SUB = subject.

b. ʔu–gʷəč'–t čəɬ ti sqʷəbayʔ
PFV–seek–ICS 1PL.SUB DEF dog
'we sought the dog'

c. ʔu–gʷəč'–t čəxʷ ti sqʷəbayʔ
PFV–seek–ICS 2SG.SUB DEF dog
'you sought the dog'

d. ʔu–gʷəč'–t čələp ti sqʷəbayʔ
PFV–seek–ICS 2PL.SUB DEF dog
'you guys sought the dog'

e. ʔu–gʷəč'–t Ø ti sqʷəbayʔ
PFV–seek–ICS 3SUB DEF dog
'he/she/it/they sought the dog' (Hess, 1995: 10)

Pronominal subjects are marked by one of a series of matrix subject clitics, the third person in this series being Ø and not making a distinction for number (Beck, 2000). A peculiarity of Lushootseed, shared to a certain extent by some other languages in the family (Gerdts, 1988; Kinkade, 1990), is that transitive sentences with both an NP subject and an NP object are disallowed (Hess, 1973). Sentences with third-person subject and object undergo obligatory pronominalization of the subject, surfacing as sentences like (1e). An interpretation of (1e) where the NP following the verb is the subject/AGENT (i.e., 'the dog sought him/her/it/them') is disallowed by what Gerdts (1988) refers to as the One-Nominal Interpretation Law. Interpretation of transitive clauses in discourse is facilitated by a reference-tracking system built around a strong constraint that subjects be topical (Beck, 2000; Kinkade, 1990).

In contexts where the identity of the subject of a transitive clause is not recoverable from discourse, or where both event-participants must be specified for communicative reasons, the passive voice is used:

(2) ʔu–gʷəč'–t–b ʔə ti č'ač'as ti sqʷəbayʔ
PFV–seek–ICS–PASS PR DEF child DEF dog
'the dog was sought by the boy'[2] (Hess, 1995: 23, ex. 6a)

Like the English passive, the passive in Lushootseed promotes the direct object to subject (Sub) and demotes the active voice subject to an oblique agentive complement (AgCo) phrase, introduced by a preposition (*ʔə*). The order of arguments can be either Sub >> AgCo or, as shown here, AgCo >> Sub, the order in (2) being more prevalent.

One of the more remarkable characteristics of Lushootseed syntax is the flexibility it displays with respect to which parts of speech are eligible syntactic predicates (Beck, 2002). As in most Salishan languages (Kinkade, 1983), words corresponding to English verbs, nouns, adjectives, adverbs, and demonstrative pronouns are all potential predicates, as in the nominally-predicated expressions in (3):

(3) a. ʔaciɬtalbixʷ čəd
Indian 1SG.SUB
'I am an Indian' (Hess & Hilbert, 1976: vol. I, 36)

b. sʔuladxʷ tiʔiɬ
salmon DIST
'that is a salmon' (Hess & Hilbert, 1976: vol. I, 7)

Non-verbal predicates like *ʔaciɬtalbixʷ* 'Indian' in (3a) take the same subject inflections as do the verbal predicates in (1) above; copular constructions with NP or demonstrative subjects like that in (3b) simply juxtapose subject and predicate, the latter appearing in sentence-initial position. Unlike nouns and other

[2] The translation "boy" is from *č'ač'as* 'child' and the non-feminine determiner, *ti*; 'girl' would be *tsi č'ač'as*.

nominal elements used as arguments, nominal predicates do not take determiners unless the sentence identifies the subject with a specific individual.

Non-verbal predicates are not confined to simple expressions of identity like those in (3); constructions like that in (4), with a nominal predicate and a complex nominal acting as subject, are quite routine:

(4) wiw'su ti?ə? ?u–čala–d ti?ə? sq^{w}əbay?
children PROX PFV–chase–ICS PROX dog
'those who chased the dog are the children' (Hess, 1995: 99)

The syntactic predicate in (4) is the bare noun *wiw'su* 'children', while the subject is a relative clause headed by the determiner *ti?ə?* (Beck, 2002). It is the predicative use of nouns in constructions like those in (3) and (4) which have led some researchers (e.g., Jelinek & Demers, 1994; Kinkade, 1983; Kuipers, 1968) to argue that there is no distinction between nouns and verbs in Salishan languages. While this is probably an over-reaction,[3] it is the case that part of speech plays less of a role in the selection of the syntactic predicate of a sentence than it does in most other languages, where there is a strong tendency for sentences to have verbal predicates. Instead, predicate selection in Lushootseed depends crucially on Communicative Structure.

2 Effects of Thematicity on syntactic structure

The component of Sem-CommS that has the greatest effect on Lushootseed syntax is Thematicity, which is the driving force behind the selection of the syntactic predicate. A similar effect is discussed for another Salishan language, Nuxalk (Bella Coola), in Davis & Saunders (1978). In Davis & Saunders' terms, the structure of a Nuxalk clause involves a bi-partition between the part of the sentence that is the "Comment" and that which is "Topic." In our terms, this translates into the Nuxalk sentence being organized so that the Sem Rheme is expressed as the syntactic predicate and the Sem Theme as its subject. Beck (1997) argues that Lushootseed has the same pattern, as shown by question-and-answer pairs such as that in (5):

(5) a. ?u–?əx̌id k^{w}i ki–kəwič
PFV–what.happen REM ATTN–hunchback
'what happened to Little Hunchback?'

— ?u ƛ'al' bə=?u–sax^{w}əb–dx^{w}–but ti?iɬ ki–kəwič
INTJ also ADD=PFV–run–DC–REFL DIST ATTN–hunchback
'oh, Little Hunchback also managed to escape' [DM Basket Ogress, lines 79–80][4]

The question in (5a) is a narratively-focused question asking about an event in which a particular Thematic event-participant is involved, and elicits a narratively-focused response with a Rhematic verbal predicate. The question in (5), however, asks for the identity of an unknown participant in a Thematic event, this event being expressed as a headless relative clause in subject position of a sentence whose predicate is the interrogative word *stab* 'what?'. The response mirrors this structure exactly, substituting the requested information for the interrogative, giving us a sentence with a Rhematic nominal predicate.

Subsequent work has found that this pattern also occurs more generally in Lushootseed narrative and other discourse contexts where the event is Thematic (and, generally, Given) and the Rheme is an event-participant or some other non-verbal element of the sentence (Beck, 2000; Beck, 2002). This is an interesting situation from a theoretical point of view in that traditional approaches to syntax generally approach clause structure as being built around phrasal projections of lexical elements whose part of speech (or the projections of the functional/inflection categories associated therewith) determine whether particular elements are realized in predicate or argument positions. Thematicity in such approaches tends to be

[3] The position that Salishan languages do not distinguish nouns and verbs is not the current consensus position held by most specialists in these languages—cf. van Eijk & Hess (1986), Kroeber (1999), and Beck (2002).

[4] Materials not cited as being from published sources are drawn from unpublished texts kindly provided by Thom Hess; these are referred to by speaker's initial followed by title of the text and line number(s).

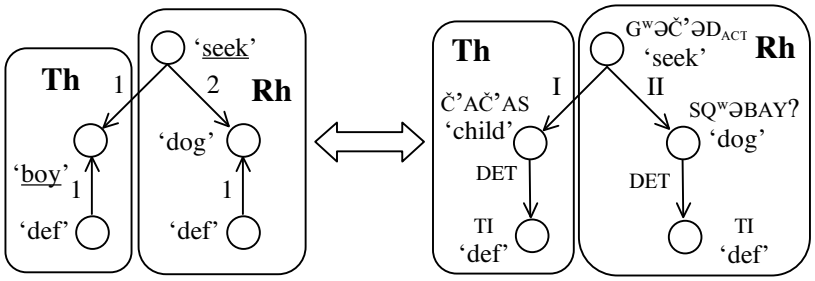


Figure 1. *ʔugʷəč'əd* (*ti č'ač'as*) *ti sqʷəbayʔ* '(the boy) sought the dog'

treated as a secondary phenomenon with less import for clause structure; however, data like that in (5) shows that Thematicity plays a much more fundamental role in Salishan languages, and suggests that it might be worthwhile to re-examine some of these traditional assumptions. In the following section, I will illustrate how the effects of Thematicity on syntactic structure can be modeled from the perspective of text generation, using some of the formalisms from Meaning-Text Theory, beginning with the selection of the syntactic predicate (Section 2.1), and then moving on to constraints on grammatical voice (2.2).

2.1 Constraints governing the selection of syntactic predicates

From the point of view of text synthesis, the initial step in mapping between SemR and DSyntR is the selection of what is called the *entry node*, that semanteme (or configuration of semantemes) in the SemR that will be lexicalized as the top node of the DSyntS (the matrix predicate). In the SemR in Figure 1 above, for example, the entry node is 'seek', which is lexicalized as the verb *gʷəčəd*, the matrix predicate in the DSyntR. One of the reasons that 'seek' is chosen is that it is the *Communicatively Dominant Node* in its sub-network of the SemR. In this context, Communicatively Dominant means that the sub-network of the SemS 'past' → 'seek' → 'dog' can be reduced to 'seek' without changing its referent—that is, its referent is a seeking event rather than, say, a dog (if 'dog' were the Comm-dominant node in this configuration, the corresponding DSynt tree would be a relative clause, 'the dog that was sought').[5]

The second consideration that determines the selection of 'seek' as the entry node is the part of speech with which its meaning is lexicalized. In English, 'seek' corresponds to the meaning of a verb, LOOK(FOR), which is selected as the top node of the DSyntS of *the boy sought the dog*. In most languages, part of speech is the primary factor governing the selection of entry nodes: by far the majority of the world's languages select as entry node semantemes that are semantic predicates and whose most natural expression is a verb.[6] However, Lushootseed departs from this pattern in preferring elements that are Rhematic, whatever their lexical class, allowing for a wide range of non-verbal predicates.

Procedural rules for the selection of entry nodes in languages like English, Russian, and French are set out in work by Iordanskaja and Polguère (Iordanskaja, 1990; Iordanskaja & Polguère, 1988) and discussed in the context of Sem-CommS by Mel'čuk (2001). Without getting bogged down in technical details, the relevant rules proposed by these authors can be restated as the ranked constraints in (6):

(6) Semantically-Predicative Entry Node (SPEN)
The entry node is a semantic predicate.

Verbal Entry Node (VEN)
The entry node is most naturally lexicalized as a verb.

[5] A more detailed discussion of Comm-Dominance and methods for determining the Comm-Dominant node can be found in Iordanskaja and Polguère (1988) and Mel'čuk (2001: 29ff.).

[6] This, of course, may depend on language-specific characteristics of the lexicon such as the (non-)existence of verbal expressions for certain semantemes (the case in point being the absence of a verbal expression of 'be' in Lushootseed—see below). The term "most naturally" is deliberately vague, allowing for lexical and other types of idiosyncrasies to override more rigid considerations such as the existence in the lexicon of a direct expression of a particular semantic predicate (*viz.* the case of the English expression BE HUNGRY which is the most common expression of the stative predicate 'hunger', in spite of the existence of the verb HUNGER).

Comm-Dominant Entry Node (CDEN)
The entry node is the Comm-dominant node in its sub-network of the SemR.

Rhematic Entry Node (REN)
The entry node is a part of the Sem Rheme.

In most languages, these constraints are ranked in the order given here, expressing the near-universal preference for verbal syntactic predicates:

(7) SPEN, VEN >> CDEN >> REN

In Lushootseed, on the other hand, the same four constraints are also present, but they are ordered differently, as in (8):

(8) REN >> CDEN >> SPEN, VEN

These rankings express the Lushootseed preference for Rhematic syntactic predicates, irrespective of their lexical class. The differences in these two ranking systems account neatly for the differences in the selection of syntactic predicate seen between English and Lushootseed.

The most straightforward case, a narratively-focused sentence with a verbal predicate such as (1e), is that shown in Figure 1 above. This sentence represents a common case in narratives, which are typically structured around one or more Topical participants. Topical participants tend to be expressed at the sentence-level as Themes, whose actions—and (to a lesser extent) the other participants with which they interact—tend to be Rhematic at the sentence-level. Thus, the SemR in Figure 1 is partitioned into two SemComm areas, the Theme containing the topical event-participant, 'child', and the Rheme containing the semantic predicate 'seek' and its SemA 2 'dog'. This sentence is an appropriate response to a question such as "What did the boy do?" or could be part of a story relating the boy's activities.

In both English and Lushootseed, the SemR in Figure 1 is expressed as an ordinary verbally-predicated sentence, given that the node 'seek' satisfies all four constraints for being the top node in DSyntR—that is, it is a semantic predicate, it is naturally expressed as a verb, it is a Comm-Dominant node (indicated by underlining), and it belongs to the Sem Rheme. This is not the case for the other nodes in the SemR, as can be seen in Table 1, which presents—using an Optimality-Theory (Prince & Smolensky, 2004) style tableau—all five candidate entry nodes in the SemR in Figure 1 as evaluated against the four constraints, ordered as per the Lushootseed rankings shown in (8).

Candidate Node	SPEN	VEN	CDEN	REN
☞ 'seek'				
'child'	*!	*		*
'dog'	*!	*	*	
'def'$_{Rh}$		*!	*	
'def'$_{Th}$		*!	*	*

Table 1. Lushootseed constraint rankings for Figure 1

Here, 'seek' is the superior candidate as it incurs no constraint violations whatsoever, whereas the remainder of the candidates violate at least two of the constraints. Because 'seek' satisfies all four constraints, the difference in rankings between (7) and (8) are of no consequence, and English and Lushootseed syntacticize this SemR in essentially the same way (although Lushootseed requires pronominalization of the NP subject in SSyntR, as noted in Section 1).

For sentences with non-verbal predicates such as that in (3b), shown in Figure 2 below, we see the same principles at work, although here the difference in constraint rankings results in very different structures. This sentence, a felicitous response to the question "what is that?", is straightforward in propositional terms, its Sem S consisting of a semantic predicate 'be' with two Semantic actants (SemAs), indi-

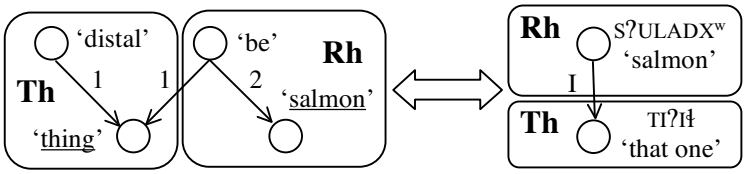


Figure 2. *sʔuladxʷ tiʔiɬ* 'that is a salmon'

cating the identification of a particular entity pointed to by the speaker (SemA 1)—as a member of a designated class of entities (SemA 2).

In languages with a copula, the semanteme 'be' would be realized as that language's equivalent of the English lexeme BE. Even though 'be' is not a Communicatively-Dominant mode in the SemR (i.e., the utterance is an assertion of "salmonhood" rather than of "being"), the constraint-rankings given in (7) guarantee that in English it will be realized as the syntactic predicate, giving us the English sentence *this is a salmon*. Thus, as shown in Table 2, although 'be' violates the second-ranked constraint CDEN, it satisfies SPEN, whereas the nodes 'salmon' and 'thing' do not. The next-best candidate, 'distal', is not part of the Sem Rheme and does not have a natural verbal expression, and therefore is also excluded.

Candidate Node	SPEN	VEN	CDEN	REN
☞ 'be'			*	
'salmon'	*!	*		
'thing'	*!	*		*
'distal'		*	*	*!

Table 2. English constraint rankings for Figure 2

The different rankings Lushootseed accords to the constraints in (6) result in the selection of a different entry node, shown in Table 3.

Candidate Node	REN	CDEN	SPEN	VEN
☞ 'salmon'			*	*
'be'		*!		*
'thing'	*!		*	*
'distal'	*!	*		*

Table 3. Lushootseed constraint rankings for Figure 2

As in English, the fact that 'be' is not the Communicatively-Dominant node of the Sem Rheme triggers a violation of CDEN; however, this constraint is ranked more highly that SPEN, and so 'be' is rejected as a candidate. The non-Rhematic nodes, 'thing' and 'distal', are eliminated by the highest-ranked constraint, REN. It should also be pointed out that in Lushootseed there is no equivalent of BE and, hence, no lexical element to appear in the DSyntS. Instead, SemA 2 of 'be', 'salmon', becomes the DSynt predicate and takes SemA 1 as its first Deep-Syntactic actant (DSyntA I), the semanteme 'be' finding its expression in the construction NPRED → N rather than in a lexical item, as it does in most languages.

In terms of its Sem-CommS, the sentence in Figure 2 can be compared with the sentence in (9), whose syntactic predicate is the proximal demonstrative *tiʔəʔ* and which would be the answer to a question such as "which stone is it?":

(9) tiʔəʔ tə čʼǎʼaʔ
PROX NDEM stone
'the stone is this one' (Hess, 1995: 81, ex. 5)

The SemR for this sentence is shown in Figure 3 below. Unlike Figure 2, however, the DSyntR in Figure 5 takes the Rhematic SemA 1 of 'be' as its syntactic predicate rather than SemA 2—in other words, the

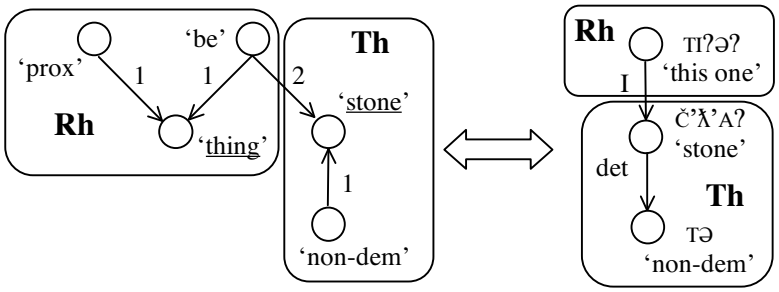


Figure 3. *tiʔəʔ tə čʼǎʼaʔ* 'the stone is this one'

entity whose identity is established by the predication becomes the DSynt-predicate, rather than the identity attributed to it, as in Figure 2. The source of this difference lies in the Sem-CommS. In Figure 2, SemA 1 (the entity being identified) belongs to the Sem Theme and SemA 2 (the identity attributed to it) lies within the Sem-Rheme. The fact that Lushootseed requires that the syntactic predicate belong to the Rheme results in SemA 2 surfacing as the DSynt predicate in Figure 2 and Sem A 1 becoming predicate in Figure 3—the Sem-CommS of the sentence, rather than its propositional content, determining DSyntS.

The contrast in constraint-rankings quite easily accounts for these differences. The most natural English sentence expressing this SemR is something like ***this*** *is the stone*, where the entry node continues to be the semantic predicate 'be'. This outcome is predicted by Table 4.

Candidate Node	SPEN	VEN	CDEN	REN
☞ 'be'			*	
'stone'	*!	*		*
'thing'	*!	*		
'proximal'		*!	*	
'non-demonstrative'		*!	*	*

Table 4. English constraint rankings for Figure 3

Once again, although 'be' is not the Communicatively Dominant node, it is selected as entry node because it is a semantic predicate and because its most natural expression is verbal. 'stone' and 'thing' are not semantic predicates, whereas the semantic predicates 'proximal' and 'non-demonstrative' have no natural verbal expression in English. For Lushootseed, the relatively low-ranking of the constraints SPEN and VEN selects 'thing' as the entry node as in Table 5:

Candidate Node	REN	CDEN	SPEN	VEN
☞ 'thing'			*	*
'stone'	*!		*	*
'be'		*!		*
'proximal'		*!		*
'non-demonstrative'	*!	*!		*

Table 5. Lushootseed constraint rankings for Figure 3

As a result, the syntactic predicate becomes *tiʔəʔ* 'this one', lexicalizing the configuration of semantemes 'prox' → 'thing' as a demonstrative.

The same principles come into play in expressions like (4), shown in Figure 4, which answers the question "who chased the dog?". The English constraint-rankings for Figure 4 are shown in Table 6:

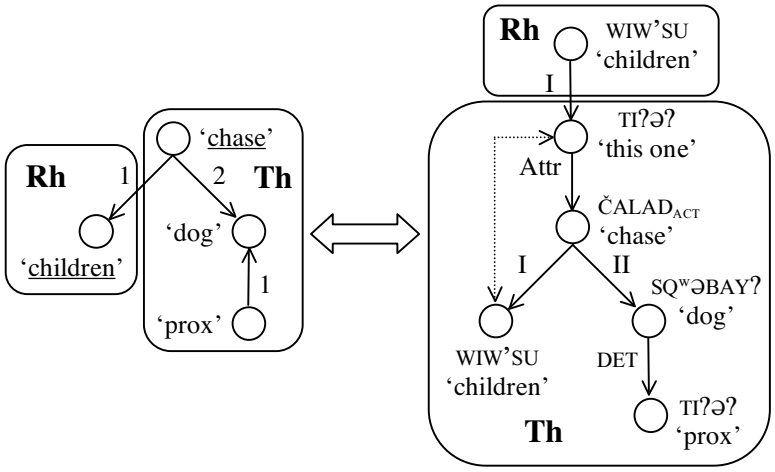


Figure 4. *wiw'su tiʔəʔ ʔučalad tiʔəʔ sqʷəbayʔ* 'those who chased the dog are the children'

Candidate Node	SPEN	VEN	CDEN	REN
☞ 'chase'				*
'children'	*!	*		*
'dog'	*!	*	*	*
'proximal'		*!	*	*

Table 6. English constraint-rankings for Figure 4

In this case, the high ranking of SPEN and VEN ensure that 'chase', which has a lexical equivalent in the verb CHASE, is selected as the entry node, giving us ***the children*** *chased the dog*, the marked prosody indicating that, exceptionally, the Thematic element is not Subject.

As shown in Table 7, the Lushootseed treatment of the same SemR gives different results:

Candidate Node	REN	CDEN	SPEN	VEN
☞ 'children'			*	*
'chase'	*!			
'dog'	*!	*	*	*
'proximal'	*!	*!		*

Table 7. Lushootseed constraint-rankings for Figure 4

Here, there is an eligible semantic predicate in the SemS, 'chase' which is most naturally lexicalized as a verb, ČALAD 'chase'; however, 'chase' is Thematic rather than Rhematic, and so the constraint rankings select the node 'children' rather than the node 'chase', reflecting the relatively low-ranking of the part-of-speech related constraints SPEN and VEN. The high precedence Lushootseed gives to the Sem-Comm status of elements thus leads to the creation of the nominally-predicated sentence *wiw'su tiʔəʔ ʔučalad tiʔəʔ sqʷəbayʔ* 'those who chase the dog [are] the children'. The form of the headless relative clause that appears as DSyntA I (that is, as the argument that will become subject in SSyntR) is the result of the interaction of further sets of constraints governing syntacticization (specifically, constraints requiring the realization of DSyntAs of verbs as nouns or nominal expressions).[7]

2.2 Thematicity and voice

Lushootseed makes use of passive voice in a way that parallels its function in many of the world's languages, that of ensuring the alignment of sentence-level Theme (and, by extension, discourse Topic) and

[7] The elision of the co-referential NPs is part of pronominalization in the DSyntR ⇔ SSyntR component.

the syntactic relation Subject (Keenan, 1976; Li & Thompson, 1976). As in most languages, the primary Communicative contrast between an active sentence such as that in (1e) and a passive sentence such as that in (2) has to do with the Thematicity of the SemA which is realized as DSyntA I: in Lushootseed, DSyntA I must be an expression of a SemA contained in the Sem Theme. In (1e), the Thematic SemA is SemA 1, 'child', giving us the DSyntR shown in Figure 1; however, in (2) (a felicitous answer to the question "what happened to the dog?" or part of a discourse episode centred on the dog), the Theme is SemA 2, the PATIENT 'dog'. In such a context, a sentence such as (1e) in the active voice is ruled out by the constraint given in (10):

(10) Thematic DSyntA I (ThDAI)
DSyntA I of the main predicate must be Thematic.

This constraint is a formalization of the strong tendency in natural languages for subjects to be Thematic and Topical. In Lushootseed, it requires the use of the passive *čalatəb* 'be sought', whose government pattern assigns the expression of the SemA 2 of 'seek' to the role of DSyntA I rather than assigning this role to SemA 1 as in the active voice.

Because of the central role that Thematicity plays in Salishan reference-tracking, the constraint in (11) is virtually inviolable for Lushootseed, the only higher-ranked constraint being an essentially morphological restriction on passives which blocks the realization of first- and second-person agentive complements (Hukari, 1976; Jelinek & Demers, 1983):

(11) 3rd Person Passive AGENT (3PassAgt)
The DSynt II of a passive verb must be third-person.

This makes the passive in Lushootseed a much more straightforward proposition than it is in, say, English or many other familiar languages where it is more closely tied to textual, stylistic, and discourse-level considerations which are related to, but potentially independent of, Sem-CommS proper (see, for example, Givón, 1994; Shibatani, 1988).

3 Conclusion

As the preceding discussion illustrates, the Communicative Structure of Lushootseed sentences is one of the primary determinants of their syntactic structure. This is seen most dramatically in the constraints governing the selection of the syntactic predicate, which in Lushootseed and many other Salishan languages depends on Thematicity rather than part of speech. The priority placed on Thematicity in this context offers a striking typological contrast to the majority of the world's languages, which place a much higher priority on considerations of semantics and part of speech in the selection of syntactic predicates. Approaching Lushootseed clause structure from the point of view of Communicative Structure also has the advantage of resolving some long-standing debates in Salishan studies concerning the (non-)existence of the distinction between nouns and verbs, allowing us to keep this useful (and probably universal) distinction while at the same time accounting for the unique properties of Salishan syntax.

Acknowledgements

This work owes a great deal to many people who have had a hand in its progression over the years. Pride of place goes of course, to Igor Mel'čuk, who provided the theoretical framework guided my understanding of it, and Thom Hess, who did the same for the data. I would also like to acknowledge the many Lushootseed speakers, now passed, who granted us a window into their language.

References

Beck, David. 1997. Theme, Rheme, and Communicative Structure in Lushootseed and Bella Coola. In Leo Wanner (ed.), *Recent trends in Meaning-Text Theory*, 93–135. Amsterdam: Benjamins.

—. 2000. Semantic agents, syntactic subjects, and discourse topics: How to locate Lushootseed sentences in space and time. *Studies in Language* 24: 277–317.
—. 2002. *The Typology of Parts of Speech Systems*. New York: Routledge.
Czaykowska-Higgins, Ewa, & M. Dale Kinkade. 1998. Salish languages and linguistics. In Ewa Czaykowska-Higgins & M. Dale Kinkade (eds.), *Salish Languages and Linguistics: Theoretical and Descriptive Perspectives*, 1–68. Berlin: Mouton de Gruyter.
Davis, Philip W., & Ross Saunders. 1978. Bella Coola syntax. In Eung-Do Cook & Jonathan Kaye (eds.), *Linguistic Studies of Native Canada*, 37–65. Vancouver: UBC Press.
Gerdts, Donna B. 1988. *Object and Absolutive in Halkomelem Salish*. New York: Garland.
Givón, Talmy. 1994. *Voice and Inversion*. Amsterdam: Benjamins.
Hess, Thomas M. 1973. Agent in a Coast Salish language. *International Journal of American Linguistics* 39: 89–94.
—. 1995. *Lushootseed Reader with Introductory Grammar, Volume I*. Missoula: UMOPL.
—. 1998. *Lushootseed Reader, Volume II*. Missoula: UMOPL.
—. 2006. *Lushootseed Reader, Volume III*. Missoula: UMOPL.
Hess, Thomas M., & Vi Hilbert. 1976. *Lushootseed: An Introduction*. Seattle: University of Washington.
Hukari, Thomas E. 1976. Person in a Coast Salish Language. *International Journal of American Linguistics* 42: 305–318.
Iordanskaja, Lidija N. 1990. Ot semantičeskoj seti k glubinno-sintaksičeskomu derevu: pravila naxoždenija veršiny dereva. In Z. Saloni (ed.), *Metody formalne w opisie języków słowiańskich, Festschrift Jurij Apresjan*, 33–46. Białystok: Universytet Warszawski.
Iordanskaja, Lidija N., & Alain Polguère. 1988. Semantic processing for text generation. *Proceedings of the International Computer Science Conference '88*: 310–318.
Jelinek, Eloise, & Richard A. Demers. 1983. The agent hierarchy and voice in some Coast Salish languages. *International Journal of American Linguistics* 49: 167–85.
—. 1994. Predicates and pronominal arguments in Straits Salish. *Language* 70: 697–736.
Keenan, Edward. 1976. Toward a universal definition of "subject". In Charles N. Li (ed.), *Subject and Topic*, 303–333. New York: Academic Press.
Kinkade, M. D. 1983. Salishan evidence against the universality of "noun" and "verb". *Lingua* 60: 25–40.
—. 1990. Sorting out third persons in Salish discourse. *International Journal of American Linguistics* 56: 341–360.
Kroeber, Paul D. 1999. *The Salishan Language Family: Reconstructing Syntax*. Lincoln, NB: University of Nebraska Press.
Kuipers, Aert H. 1968. The categories verb-noun and transitive-intransitive in English and Squamish. *Lingua* 21: 620–626.
Lambrecht, Knud. 1994. *Information Structure and Sentence Form: Topic, Focus, and the Mental Representation of Discourse Referents*. Cambridge: Cambridge University Press.
Li, Charles N., & Sandra A. Thompson. 1976. Subject and topic: A new typology of language. In Charles N. Li (ed.), *Subject and Topic*, 457–489. New York: Academic Press.
Mel'čuk, Igor A. 2001. *Communicative Organization in Natural Language: The Semantic-Communicative Structure of Sentences*. Amsterdam: John Benjamins.
Prince, Alan, & Paul Smolensky. 2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Malden, MA: Blackwell.
Shibatani, Masayoshi. 1988. *Passive and Voice*. Amsterdam: Benjamins.
Vallduví, Enric. 1992. *The Informational Component*. New York: Garland.
van Eijk, Jan, & Thomas M. Hess. 1986. Noun and verb in Salishan. *Lingua* 69: 319–31.

Semantics of Attenuated Comparatives in Russian

Igor Boguslavsky
Universidad Politécnica de Madrid
Campus de Montegancedo, 28660 Boadilla del Monte, Madrid, Spain
Igor.Boguslavsky@upm.es

Leonid Iomdin
Institute for Information Transmission Problems, Russian Academy of Sciences
B. Karetnyj Per.19. Moscow, 127994, Russia
iomdin@iitp.ru

Abstract[1]

The paper discusses the semantics of Russian comparatives with the attenuative prefix *po-* like *pobol'še* 'somewhat more' or *poumnee* 'a bit cleverer'. It is shown that they have two major types of usage, the comparative one and the selective one. The former is close to the semantics of the standard comparative degree of adjectives and adverbs, while the latter is closer to the semantics of the superlative. Within the comparative type a special case is described: the correlative construction which contains two attenuatives. Its outstanding feature is cross-filling of valency slots of the comparatives.

1 Attenuated Comparatives in the Lexical System of the Language

Comparative forms of adjectives and adverbs with the prefix *po-,* like *pobol'še, poumnee* (roughly, 'somewhat more', 'a bit cleverer'), are so common in Russian and so fully documented in grammars and linguistic literature that it may raise the question of whether another study in this area could really add anything new to the existing knowledge. Indeed, such linguistic units are semantically straightforward; they all share a meaning component of attenuation close to 'somewhat' and are, at first glance, not enigmatic at all.

We will however attempt to demonstrate that these units have very nontrivial properties, some of which have not been discussed in literature and deserve close consideration.

What do we know about attenuated comparatives in Russian? The basic facts can be found in the famous Victor Vinogradov's book (1947, reissued in 1972) and all Russian academic grammars. Vinogradov discusses these units comparing them with regular comparatives:

"In forms ending with *-ee*, *-e*, *-še* the meaning of the comparative degree is not complicated by any additional shades. These forms may combine with the prefix *po-*. The prefix *po-* usually attenuates the degree of prevalence of a quality in one of the compared objects. For example: *ljudi pobednee nas* 'people somewhat poorer than we'*, učeniki pomolože* 'students (who are) a bit younger'" (Vinogradov 1972: 210-211)[2].

The most recent Russian academic grammar describes the meaning of such forms in more or less the same terms: "Uninflected words (adjectives and adverbs) which are formed from the comparative degree with the help of the prefix *po-* possess the categorical properties and the meaning of the comparative but have the general meaning of an attenuated, moderate degree of manifestation of a property." (Grammatika 1980, v.1, p. 565).

In this paper, we will focus on the semantic peculiarities and, on a lesser scale, syntactic properties of constructions that contain such units, largely disregarding the issue of their grammatical or lexicographic status. This means that we will not consider arguments for and against any concrete theoretical decisions, e.g. (a) whether or not these units belong to two different parts of speech – adjectives and adverbs, or else

[1] This study was partly supported by two grants (07-06-00339 и 08-06-00373) from the Russian Foundation for Basic Research, whose contribution is gratefully appreciated.

[2] Here and below, all translations have been made by the authors.

constitute a separate part of speech, together with the comparative or without it; (b) whether or not the paradigms of the respective adjectives and adverbs include the forms with *po-* or else the latter are their word formation derivates and constitute independent lexemes; (c) whether or not these units are comparatives in need of a grammeme of an additional grammatical category (attenuativity or the like), or else regular comparatives and *po*-comparatives should be ascribed different values of a certain more general grammatical category

Still, to avoid confusion, we will proceed from the following three assumptions:

1) There are adjectives and adverbs with *po-*. In *Prinesi kamen' potjaželee* ≈ 'bring a rather heavy stone' or 'bring a somewhat heavier stone',[3] we have a *po*-adjective, whilst in *Nado rabotat' poaktivnee* ≈ 'one has to work a bit more actively' there is a *po*-adverb. In an ambiguous sentence

(1) *On byl povyše*,

the latter word may be interpreted as an adjective if (1) means 'he was somewhat taller' (and relates to *On byl vysokim* 'he was tall') and as an adverb if (1) means 'he was situated in a somewhat higher place' (and relates to *On byl vysoko* 'he was in a high place').

2) The words with *po-* – we will henceforth refer to them as attenuatives, following Igor Mel'čuk (1998)[4] – like regular comparatives, are included into the paradigms of the respective adjectives and adverbs. If an adjective or an adverb has more than one sense, attenuatives may appear in the paradigms of all of the respective lexemes, or only in some of them. For instance, the adjectives *legkij 1* 'lightweight' and *legkij 2* 'not difficult' both have attenuatives (*voz'mi molotok polegče* 'take a slightly lighter hammer' vs. *dajte emu zadaču polegče* 'give him a slightly less difficult task', whereas the polysemous adjectival vocable *ostryj* readily generates the attenuative form in some of the senses (*nož poostree* 'a somewhat sharper knife', *sous poostree* 'a somewhat hotter sauce', and categorically rejects them in a terminological sense (**appendicit poostree* 'a somewhat more acute appendicitis').

3) All attenuatives are comparatives at the same time. The words under discussion must be assigned the comparative grammeme since they appear in many types of syntactic environment typical of comparatives, including the comparative constructions, in which the second element of comparison is expressed by the genitive or introduced with the conjunctions *čem* or *neželi* 'than', as in *On zarabatyvaet bol'še direktora* 'he earns more than the director' vs. *On zarabatyvaet pobol'še direktora* 'he earns somewhat more than the director'; *On starše, čem ego brat* 'he is older than his brother' vs. *On postarše, čem ego brat*' he is somewhat older than his brother'.

The availability of attenuatives in the paradigms of adjectives and adverbs is quite regular: most qualitative adjectives and adverbs have these forms even though some have not. Interestingly, the presence of standard comparatives is no guarantee that the word will have an attenuated form: *nužnee* 'more needful', *neobxodimee* 'more necessary', *želatel'nee* 'more desirable' exist but *??ponužnee, *??poneobxodimee, *??poželatel'nee* do not. Attenuated forms of adjectives or adverbs of a complex morphological structure (with prefixes, two radicals etc.) are formed with difficulty, even if standard comparatives are part of the corresponding paradigm: *??pobezavarijnee* 'a bit more accident-free', **poneiskrennee* 'somewhat more insincere', *??ponesvoevremennee* 'somewhat more untimely', etc. On the other hand, the authors are unaware of any cases when an adjective or adverb would have an attenuative but no standard comparative in its paradigm.

It should also be noted that attenuatives are formed precisely in the same way as standard comparatives are, with the same morphological structure (excluding the prefix, of course), not depending on the productivity or non-productivity of the morphemes used or on morphological (semantically void) variation: *rezče – porezče* 'sharper – a bit sharper', *ran'še – poran'še* 'earlier – a bit earlier', *bystree/bystrej – pobystree/pobystrej* 'faster – a bit sharper', etc.

[3] For reasons of space, we will not always provide multiple English equivalents of examples with words under discussion, even though in many cases, including (1), the respective utterances can be interpreted in more than one way depending on the particular use of such words.

[4] Igor Mel'čuk sees in these word forms an attenuative quasi-grammeme, i.e. a grammeme defined on a part of the paradigm rather than the whole paradigm.

To our best knowledge, there are no lexicographical resources, traditional or computerized, that systematically indicate presence or absence of attenuatives for Russian adjectives or adverbs. A significant step in this direction is made in the computer dictionary of the multi-purpose linguistic processor ETAP-3 (Apresjan *et al*. 1992, 2003), developed with the immediate participation of both authors. Here, the information on attenuatives is given for each lexeme of every adjective and adverb.

According to our observations,[5] attenuatives are used on an incredibly wide scale in modern Russian, especially in the colloquial speech. The meanings conveyed by these lexical units, which are explained below, appear to be very useful for Russian speakers. We observe that these forms can be acquired very early in life; a two-year-old boy, who is only starting to use basic prepositions, says *lezu ešče povyše* 'I am climbing a bit higher still'. Attenuatives are occasionally formed from practically any qualitative adjectives and adverbs, and sometimes even from relative adjectives and adverbs whose meaning normally excludes quantification. Note some illustrative examples taken from real texts of various genres:

Ego absoljutnyj slux ***poabsoljutnee*** *moego budet!* 'His absolute ear (for music) will be somewhat more absolute than mine';

Tibet i tak imeet širokuju avtonomiju, už ***poavtonomnee*** *Burjatii v sostave RF* 'Tibet already has a broad autonomy, in any case it is rather more autonomous than Buryatia within Russia';

A šumerskij-to, ja smotrju, ***poaggljutinativnee*** *tjurkskix budet* 'The Sumerian language, I presume, is rather more agglutinative than Turkic languages' etc.

Attenuatives are very frequently used in Russian translations from languages where there is nothing even remotely similar to such units, e.g. English:

Kak raz v etot moment rukovoditel'nica xora, mogučego telosloženija ženščina v tvidovom kostjume, nastavljala ix, čtoby oni ***pošire*** *raskryvali rty, kogda pojut* (D. Salinger. "For Esmé – with Love and Squalor").

In the original, the standard comparative was used: *At the moment, their choir coach, an enormous woman in tweeds, was advising them to open their mouths* ***wider*** *when they sang.*

Podobnye razgovory, po-vidimomu, neizbežno vlekut za soboj žaždu razvlečenij, i vot, v minutu slabosti, ja predložil Džordžhu vytaščit' bandžo i popytat'sja ispolnit' kakuju-nibud' pesenku ***poveselee*** (Jerome K. Jerome. "Three Men in a Boat (To Say Nothing of the Dog)").

The original text by Jerome K. Jerome does not even include a comparative: *There seemed to be a desire for something frolicsome to follow up this conversation, and in a weak moment I suggested that George should get out his banjo, and see if he could not give us a* ***comic*** *song*.

Importantly, both Russian translations have not even a trace of artificiality and are adequate to the originals.

2 Syntactic and Combinatorial Properties of Attenuatives

We mentioned above that attenuatives share important syntactic and combinatorial properties with standard comparatives. Still, there are essential differences between these two classes of units, to be briefly outlined below.

2.1 Attenuatives and the Adverbs of Degree

Attenuatives are reluctant to allow adverbs denoting high degree, including those appearing exclusively, or primarily, with the comparative: *gorazdo, namnogo, kuda* 'much'*, ves'ma* 'quite'; cf. *gorazdo sil'nee* 'much stronger', but not **gorazdo posil'nee* 'much a bit stronger', *rabotajut kuda bystree* 'work noticeably faster', but not **rabotajut kuda pobystree* 'work noticeably somewhat faster'. In contrast, adverbs and particles denoting small degree, like *nemnogo, neskol'ko* 'somewhat', *čut'* 'a tiny bit', *slegka* 'slightly' are commonly used with the standard and the attenuated comparatives alike, as in *nemnogo ran'še <poran'še >* 'somewhat earlier', *čut' glubže <poglubže>* 'a tiny bit deeper',

[5] In collecting the material, we have been using three Russian text corpora: 1) the untagged corpus of the Institute of the Russian Language in Moscow; 2) the morphologically tagged National corpus of Russian and 3) the annotated corpus of Russian SYNTAGRUS, which offers complete syntactic tagging for each sentence. The latter corpus was created by the team to which both authors belong (see Boguslavsky *et al*. 2002). The National corpus of Russian and SYNTAGRUS can be freely accessed at www.ruscorpora.ru.

Doubtless, these combinatorial peculiarities of attenuatives are of semantic nature: if *pobol'še* means 'somewhat more' than expressions like *namnogo pobol'še* 'much somewhat more' harbour a semantic contradiction.

2.2 Attenuatives and the Negation

It is harder to explain rather whimsical rules of combining attenuatives with the particle *ne* 'not'. Such combinations are perfectly normal in contrastive situations or in cases where the appropriate use of an attenuative is questioned: *On byl ne postarše, a, naoborot, pomolože svoej nevesty* 'He was not somewhat older but, conversely, somewhat younger than his fiancée', *On k tomu vremeni budet ne postarše, a drevnim starcem s sedinami do pola* 'By that time, he will not be a bit older but an ancient old man with grey hair reaching to the floor'. On the contrary, regular negation whose scope could be confined to the attenuative, does not occur: *On ne vyše šestiletnego rebenka* 'He is no taller than a six-year-old child' but not **On ne povyše šestiletnego rebenka* 'He is not a bit taller than a six-year-old child'. A probable cause may be the communicative structure of the attenuatives, which seem to contain a conjunction of two predicates; *pobol'še X-a* means 'more than X and a bit more than X': such a conjunction does not allow a natural negation. The picture appears to be similar to the one that exists in construction like *On rezko zatormozil* ≈ 'he slammed on the brakes' which, roughly, means 'He used the brakes, and used them abruptly' and does not allow grammatical negation: **On ne rezko zatormozil* is unacceptable (Paducheva 1974, Boguslavsky 1985).

Quite expectantly, attenuatives reject, even more strongly, the negation accompanied by intensifiers: **On niskol'ko <ničut, ni kapel'ki> ne povyše šestiletnego rebenka* ≈ 'He is not a tiniest bit taller than a six-year-old child'.

2.3 Absence of Analytical Superlative and Comparative

The analytical superlative easily formed from the standard comparative, as in *Lučše vsego on umeet igrat' v futbol* 'He can play football best of all' or *On sil'nee vsex* 'He is stronger than everyone else' is strictly banned for the attenuative: **Polučše vsego on umeet igrat' v futbol* 'He can play football somewhat best of all', **On posil'nee vsex* 'He is somewhat stronger than everyone else'. Even this ban, however, is sometimes violated in colloquial speech if additional factors enter the play. A notable case is colloquial style making use of the verb *budet* (future tense of *be*) in a special sense of assumptive evaluation, very typical in attenuated contexts. To give an example, a sentence like *Brazilija posil'nee vsex budet* 'Brazil will surely be stronger than all' is likely to appear in the discussion of a football championship and perfectly grammatical. We believe that in such contexts the attenuative loses the semantic element 'somewhat' and behaves as a synonym of the regular comparative.

Let us also note for completeness that, unlike the regular comparative which has both a synthetic and an analytical variant (*čašče – bolee často* 'oftener – more often'), the attenuative can only be synthetic (*pointeresnee* 'more interesting' but not **pobolee interesnyj*.)

2.4 Syntactic Features

It is known that many Russian adjectives and adverbs have certain nontrivial syntactic features of the following nature: occupying the position of the predicate (maybe, in combination with a copula verb) they may accept subjects expressed by an infinitive or a certain type of subordinate clause (formed by conjunctions *čto* 'that', *čtoby* 'so that', an indirect question, and the like,[6] cf.

(2) *Trudno* [=predinf] *bylo poverit', čto on prav* 'It was difficult to believe that he was right';

(3) *Interesno*, [=predchto] *čto oni ne pomnjat, kto predostavil im nezavisimost'* 'It is interesting that they do not remember who gave them their independence';

(4) *Važno* [=predchtoby], *čtoby emu rasskazali pravdu* 'It is important that he should be told the truth'.

[6] For details on syntactic features like "predinf", "predchto" etc. see Iomdin-Mel'čuk-Pertsov 1975 and Apresjan *et al* 1992.

These features are shared by standard comparatives, even though they are not manifested on a similarly regular basis and may require that additional conditions be satisfied; compare (3)-(4) and

(3a) *Ešče interesnee*, *chto oni ne pomnjat, kto predostavil im nezavisimost'* 'It is even more interesting that they do not understand who gave them their independence'*;*

(4a) *Gorazdo važnee*, *čtoby emu rasskazali pravdu* 'It is much more important that he should be told the truth'.

In principle, attenuatives retain these features, too. However, they seldom come into play and the conditions imposed on acceptable utterances are even more rigid and less formalizable:

(3b) *Poverit', čto on prav, bylo nemnogo potrudnee* 'It was somewhat more difficult to believe that he was right'*;*

(4b) *Pointeresnee, požaluj, drugoe – chto oni etogo ne pomnjat.* 'Maybe another thing is a bit more interesting: that they do not remember this'.

A quite expectable conclusion could be drawn from the above. Any linguistic phenomenon manifests itself the most strongly in canonical cases and becomes washed-out, less and less expressed as one proceeds from the center to the periphery. With regard to these syntactic features, attenuatives are remote periphery.

2.5 Attenuatives as Attributes

In contrast to the above, attributive constructions to be discussed in this subsection are not peripheral with regard to the attenuatives but, conversely, are especially characteristic of them. In these constructions, the attenuated adjective is a (predominantly postpositive) modifier of a noun: *pozovite parnej pokrepče* 'call rather more strong guys'*, ja znal načal'nikov i poumnee* 'I used to know bosses who were a bit more clever' etc. Regular synthetic comparatives are unwelcome in such constructions and should be replaced by analytical ones: *Pozovite bolee krepkix parnej,* but not **Pozovite parnej krepče* или **Pozovite krepče parnej*. (Comparative attributes become more acceptable if they have their own syntactic dependents: *Ja znal načal'nikov i umnee našego* 'I knew bosses who were more clever than ours').

Of special interest is a subclass of attributive constructions formed by attenuatives where the modified word is one of the interrogative pronouns *kto* 'who' and *čto* 'what': *Pozovi kogo pokrepče* 'Call someone rather more strong' *Pojdeš' v magazin, kupi čego povkusnee* 'If you go shopping, buy something rather more delicious'.

This subclass is further represented by constructions with the interrogative adjective *kakoj* 'which, what kind' as in *Ja predlagal komandiru zakopat' granatomet v zemlju, no on prikazal tol'ko zavalit' ego kamnjami, vybiraja kakie potjaželee* 'I proposed to the commander to dig the hand-mortar but he only ordered to heap stones on it, choosing very heavy ones' (Viktor Suvorov, The Aquarium).

As a matter of fact, the constructions with pronominal adverbs like *Ego nado otpravit' kuda podal'še* 'He has to be sent to some rather remote place', *Neužheli v Pitere ili gde pobliže ne nashlos' podxodjaščego aktera?* 'Can it be that a suitable actor was not to be found in Petersburg or at some closer place' also belong to this class, even though the attenuatives occurring in the last two sentences (*podal'še, pobliže)* are formed from adverbs and not from adjectives,

In all examples of the last two paragraphs, the interrogative pronouns in fact behave as indefinite ones: *kupi č ego povkusnee = kupi č ego-nibud' povkusnee, otpravit' kuda podal'še = otpravit' kuda-nibud' podal'še.* Accordingly, these examples enrich the class of cases where interrogative pronouns are used instead of indefinite ones and, quite unexpectedly, come to resemble minor type constructions like *kto popalo* 'anybody who comes first', *gde xočeš'* 'wherever you wish' etc.[7]

[7] A detailed description of such cases can be found in Iomdin 2007.

Curiously enough, such constructions are opposed syntactically and – in all appearance – semantically to another class of cases where the interrogative pronoun is a subject or object of the predicate of a clausal complement, where this predicate is expressed by an attenuative. Compare, for instance, sentences (5a) and (6a), on the on hand, and (5b) and (6b), on the other hand.

(5a) *On vybiral kogo pokrepče* 'He was choosing the one (or the ones) rather more strong'.

(5b) *On vybiral, kto pokrepče* 'He was choosing who was rather more strong'.

In (5a), *kogo pokrepče* is a noun phrase where the attenuative is an attribute of the pronoun, whereas in (5b) *kto pokrepče* is a clause with a predicate *pokrepče* and a nominal subject *kto*.

(6a) *Naden' čto poxuže* 'Put on something rather less smart'.

(6b) *Ul'jaša, verno, rassprašivaet sejčas tam etu baryšnju, nadevaja čhto poxuže, čtoby potom zanjat'sja raskladkoj veščej* 'Ulyasha is probably questioning the young lady there, putting on some clothes that happen to be less smart in order to start arranging things later' (Boris Pasternak, The Childhood of Liuvers).

In (6a), *čto poxuže* is a noun phrase in the accusative, whilst in (6b), just like in (5b), it is a clause with a predicate *poxuže* and a nominal subject *čto*.

3 Semantics of Attenuatives

3.1 Comparative Type

Attenuative *po*-constructions have two quite different semantic varieties – the comparative one and the selective one. We will examine them in sections 3.1 and 3.2, respectively.
Po-forms belong to the comparative type if they convey the idea of comparison of two entities which possess property P in different degrees, cf.:

(7) *Tol'ko odno uxo, požaluj, podlinnee drugogo* 'one ear is perhaps a bit longer than the other'.

In (7), both elements of comparison are represented: the one that is being compared with something ('one ear') and the one with which the comparison is made ('the other ear'). We will term these elements as "first comparate" and "second comparate", respectively. In the canonical case the first comparate syntactically subordinates the comparative word, which in its turn subordinates the second comparate.

Of all constructions with attenuatives, sentences like (7) are closest of all to standard comparative constructions with non-attenuative forms of adjectives and adverbs (COMPAR). The latter can always replace the attenuative forms without any significant shift of the meaning.

Syntactic Optionality of the Second Comparate

In attenuative constructions, the second comparate is not syntactically obligatory. According to our estimates, sentences without an explicit second comparate even prevail over the ones that contain it. While regular comparatives do not necessarily require an explicit second comparate, either, attenuatives are much more eager to skip it. In (8), the *po*-form modifies the noun: (8a) has a simple modifier and (8b) a dangling one:

(8a) *Posle etogo iz ščeli vybežal tarakan pomen'še* 'After that, a smaller cockroach scurried out of the crack'.

(8b) *Iz ščeli vybežal vtoroj tarakan, pomen'še* 'A second cockroach, a smaller one, scurried out of the crack'.

The second comparate is not mentioned in either sentence but is given in the preceding context. In both cases, the COMPAR forms appear to be impossible without an explicit second comparate: **Iz ščeli vybežal tarakan men'še* and **Iz ščeli vybežal vtoroj tarakan, men'še* are both wrong. The introduction of the second comparate makes the sentence acceptable: *Iz ščeli vybežal vtoroj tarakan, men'še pervogo* 'A second cockroach, smaller than the first one, scurried out of the crack'. This requirement is dropped if COMPAR makes part of a dangled phrase, as in (9), or acts as the predicate, together with a copula verb, or without it, as in (10):

(9) *Na nem byl takoj že remen', no nemnogo uže* 'his belt was similar, but somewhat narrower'.

(10) *I jazyk počti takoj že, kak u nas, xotja i xuže* 'Their language is nearly the same as ours, although (it is) worse' (M.Shcherbakov, The Duo).

If COMPAR acts as an adverbial modifier of a verb, this restriction does not apply: *Iz ščeli vybežal vtoroj tarakan, men'še napominajuščij golodnogo volka* 'A second cockroach, less resembling a hungry wolf, scurried out of the crack'.

Semantic Obligatoriness

As we showed above, the valency slot of an attenuative corresponding to the second comparate is syntactically optional. It is important, however, that it is semantically obligatory. We term a slot semantically obligatory if it must be filled in the semantic structure even if it is not instantiated in the syntactic structure.

Consider the sentence

(11a) *Imenie svoe on prodal i kupil drugoe* 'He sold his estate and bought another one'.

The price slot for both verbs 'buy' and 'sell' is syntactically optional. (11a) does not convey any information on the price; neither does it assume that such information is present in the preceding context. The situation is different with the valency slot of the attenuated comparative which corresponds to the second comparate. Even when it is not explicitly filled with a dependent phrase in the syntactic structure, the semantic structure always contains the meaning which instantiates this slot. Sometimes, this information is given in the same sentence:

(11b) *Imenie svoe on prodal i kupil drugoe, pomen'še* 'He sold his estate and bought another one, somewhat smaller'.

In (11b) we obviously know with what the new estate is compared: it is smaller than the estate he had before. In other cases, the meaning which fills this slot is found in the preceding context:

(12) *Rjadom s lestnicej byl xoll pomen'še* 'near the stairs there was a somewhat smaller hall'.

Although (12) mentions only one hall, if the text is coherent, another one had to be introduced before.

Finally, there are cases in which the second comparate is not even mentioned in the context, but even then it makes part of the semantic structure of the sentence:

(13a) *Esli by ja znala, čto u tebja podnjalas' temperatura, ja prišla by poran'še* 'If I had known that your temperature got high, I would have come earlier' (= '...earlier than I really came ').

(13b) *Esli u tebja podnimetsja temperatura, ja pridu poran'še* 'If your temperature gets high, I will come earlier' (= '...earlier than I would come otherwise').

For all three cases (12), (13a) and (13b), it is obvious that the second comparate is not absent from the sentence in the same way as the information on the price is absent from sentences (11). The speaker always reads a definite content into this slot, while the addressee uses the context to reconstruct it. [8]

Special Case: The Correlative Construction

As we saw above, if the second comparate is not directly subordinated by the comparative word and is to be found in the context, it should be the preceding context. However, there exists a class of syntactic constructions which violates this rule: the *po*-form does not subordinate the second comparate, which is located to the right of the comparative word. These are correlative constructions.

A correlative construction is a two-part coordinative construction, such that:

(a) its both parts describe the same situation P, in which two participants are singled out – A and B. For example, in

(14) *Mal'čiki peli gromko, a devočki tixo* 'the boys were singing loudly and the girls softly',

P = 'to sing', A = 'the one who is singing', B = 'level of sound in singing';

(b) two parts of the construction denote two copies of P – P1 and P2, which differ in these participants. In (14), 'the boys' = A1, 'loudly' = B1, 'the girls' = A2, 'softly' = B2.

(c) two parts of the construction have parallel communicative structures: A1 and A2 are Themes, and B1 and B2 are Rhemes (cf. (14a)), or the other way round (14b).

(14a) *Mal'čiki* [A1=Th] *peli gromko,* [B1=Rh]*, a devočki* [A2=Th] *tixo* [B2=Rh] 'the boys [A1=Th] were singing loudly [B1=Rh] and the girls [A2=Th] (were singing) softly [B2=Rh].

(14b) *Gromko* [B1=Th] *peli mal'čiki* [A1=Rh]*, a tixo* [B2=Th] – *devočki* [A2=Rh] 'the ones who were singing loudly [B1=Th] were the boys [A1=Rh], and (the ones who were singing) softly [B2=Th] (were) the girls [A2=Rh]'.

Let us now introduce attenuatives in the rhematic components of the correlative construction:

(15) *Xrustal'nye bokaly byli pobol'še, a stekljannye pomen'še* 'the crystal wineglasses were a bit bigger, and the glass ones a bit smaller'.

Sentence (15) is remarkable in several respects. First of all, it is semantically superfluous (tautological). It contains two comparative words but they are, in a way, closed onto one another, so that we have only one act of comparison: the crystal wineglasses were a little bigger than the glass ones, and the glass wineglasses were a little smaller than the crystal ones. (15) can be paraphrased so that to leave only one comparative word, and the meaning will not change:

(15a) *Xrustal'nye bokaly byli pobol'še stekljannyx* 'the crystal wineglasses were a bit bigger than the glass ones'.

The closure of comparative words manifests itself in their having common, but criss-crossing actants. The first comparate of *pobol'še* 'a little bigger' acts as the second comparate of *pomen'še* 'a little smaller', whereas the second comparate of *pobol'še* acts as the first comparate of *pomen'še*. This phenomenon only occurs when comparative words are in the rhematic position. If the communicative structure changes, such distribution of actants is hardly possible:

(15b) *??Pobol'še byli xrustal'nye bokaly, a pomen'še stekljannye*

[8] The property of the semantic obligatoriness of the second comparate is only characteristic of the comparative use of the *po*-forms proper. Non-comparative use of these forms is discussed below, in section 3.2.

It should be noted that this phenomenon is not characteristic of standard comparatives. Sentence (16) is quite appropriate in context of (16a), but it does not imply actant criss-crossing.

(16) *Xrustal'nye bokaly byli bol'še, a stekljannye men'še* 'the crystal wineglasses were a bit bigger, and the glass ones a bit smaller'.

(16a) *Nam nužny byli bokaly emkost'ju 0,25 l, no my ix tak i ne smogli najti. Xrustal'nye bokaly byli bol'še, a stekljannye men'še* 'We needed 0.25 l wineglasses but failed to find them. The crystal wineglasses we were offered were bigger, and the glass ones were smaller'.

Sentence (16) means that the crystal wineglasses were bigger than needed and the glass ones were smaller than needed but not that the crystal wineglasses were bigger than the glass ones (even though this fact logically follows from the previous one).

The specific feature of sentences like (16) as compared with sentences which do not contain a correlative construction consists in the fact that the scope of the comparative meaning is determined univocally. Both comparates of both comparative words have fixed positions in the correlative construction without being subordinated by them.

It should be emphasized that cross-interpretation of the *po*-forms only occurs in their comparative use. Let us compare sentences (17a) and (17b), both containing the correlative construction:

(17a) *Mal'čiki peli pogromče, a devočki potiše* 'The boys were singing a little louder and the girls a little softer'.

(17b) *Mal'čiki staralis' pet' pogromče, a devočki potiše* 'The boys tried to sing rather louder, and the girls (tried to sing) rather more softly'.

The cross-interpretation of the comparative words is only natural in (17a): the boys were singing louder than the girls. Sentence (17b), in its most natural interpretation, does not compare the levels of the boys' and the girls' sound. They are characterized irrespective of one another. That means that the *po*-forms have a different type of use which will be discussed in detail in the next section.

3.2 Selective Uses

In addition to the comparative uses discussed above, attenuatives have a different type of usage, which can be illustrated by the following examples:

(18) *Tam on vybral ploščadku, gde bylo* ***pomen'še*** *komkov, sunul ladon' v zemlju i srazu oščutil, kak po pal'cam poteklo pokojnoe blagodatnoe teplo* 'He chose a piece of land where there were fewer lumps of earth, plunged his hand into the earth and felt at once how calm, life-giving warmth streamed along his fingers (N. Gladyshev, Anton's Well).

(19) *Každyj videl pered soboju smert' i staralsja tol'ko podorože prodat' svoju žizn'* 'Everyone saw death in front of him and only strived to sell his life as dearly as possible' (Nikolaj Gogol, Taras Bulba (an early version)).

Since no concrete object of comparison (second comparate) is present in such uses, they are referred to as irrelative (Vinogradov 1972 : 210-211) or absolute (Kustova 2002). We prefer to call these uses selective, since their semantic nucleus is formed by the idea of choice. (We fully agree with Galina Kustova who discussed this idea).

It is interesting to compare the selective uses of attenuatives with the comparative uses considered in the previous section. Sentence

(20) *Prinesi kamen' potjaželee* 'bring a somewhat heavier stone'

has two interpretations– (20a) and (20b):

(20a) ‘There was one stone and it proved to be too light’.

This is a comparative interpretation proper, in which (20) is synonymous with the sentence *Prinesi kamen’ potjaželee, čem etot* ‘Bring a stone somewhat heavier than this one’.

(20b) ‘No stone is within the speaker’s eyeshot yet’.

In a first approximation, it can be said that (20b) introduces a relatively heavy stone. However, this characteristic only refers to the stone itself, without any comparison with other stones. Yet, (20b) obviously considers this particular stone against the background of other stones, with which it is implicitly compared. There is a set of potential possibilities, and the stone is chosen among them. But if there is comparison, even an implicit one, then what is the role of the attenuation itself?

If we continue the quotation from Vinogradov’s book started in the beginning of this paper: “in the irrelative use of the comparative form, when concrete objects of comparison are absent, the prefix *po-* may mean ‘to the highest extent possible, with a higher degree of something than usual’. This meaning appears the most frequently in the adverbial form of the comparative degree. For example: «Ja poskoree vybralsja iz kibitki» ≈‘I got out of the caravan as soon as I could’ (Pushkin, A Journey to Arzrum)” (Vinogradov 1972: 211), we will gather that the author does not see any semantic generality between the senses introduced by *po-* in the comparative and the selective (irrelative) uses of the units considered.

Indeed, in the comparative uses proper, like (20a), the difference between the compared objects are attenuated. If we say *Etot kamen' potjaželee, chem zelenyj* ‘this stone is somewhat heavier than the green one’ it just means that the weight of this stone is not much heavier than that of the green stone. Obviously, uses like (20b) do not imply anything remotely near to the idea that this particular stone is only a little heavier than the others. Still, we believe that in such cases the prefixed comparatives convey the meaning of attenuation of a value, too. Only, this value is not that of the difference between the compared objects.

In selective uses, *po-*forms should not be confronted with the comparative degree of an adjective or adverb. Instead, they must be confronted with the superlative degree: the attenuation is observed with regard to this latter category.

Indeed, let us compare the following series of utterances:

(21a) *Prinesi samyj tjaželyj kamen'* ‘bring the heaviest stone’.

(21b) *Prinesi kamen', čem tjaželee, tem lučše <kak možno bolee tjaželyj kamen'>* ‘bring a stone, the heavier the better <bring as heavy a stone as possible>’.

(20) *Prinesi kamen’ potjaželee* ‘bring a rather more heavy stone’.

In this series of semantically similar phrases, one can easily see how the categorical character of utterances is gradually weakening. In (21a) the speaker demands that no heavier stone should exist. In (21b), strictly speaking, this is not required, but it is implied that should the heaviest stone be brought, it will suit the speaker all right. In (20) the request is weaker still. The addressee is not asked to select the heaviest stone. It will suffice if the stone makes part of the group of the heaviest stones among those available. In other words, it should be heavier than most of the stones. It is not supposed at all that the heaviest stone will better suit the speaker than, say, the second heaviest one. What is needed is that the stone should be definitely closer in weight to the heaviest stone than to the lightest one.

Thus, the semantic contribution of the attenuative is clearly different in sentences of the type (20) than in comparative constructions discussed in section 3.1. The comparative interpretation proper of the attenuative consists in transforming the meaning ‘more’ into ‘somewhat more’. In sentences of the type (20), what is attenuated is the meaning ‘more than all the others’, which turns into ‘more than most of the

others'. One can say that in comparative constructions the operator of attenuation affects the meaning of comparison, while in the selective ones it affects the meaning 'all'. Morphologically, the *po*-forms remain attenuative comparatives, while, semantically, they are attenuative superlatives.

Let us now turn to another aspect of the attenuative semantics. As mentioned above, the ideas of desire and choice are central in the meaning in the selective attenuative. Here, there is always a figure of the subject that has a number of alternatives before him and he attempts (or is impelled by someone) to select the one among them which surpasses most of the other alternatives in the degree of the property in question. Cf. the correct sentence (22a) and an unacceptable one (22b):

(22a) *On vybral jabloko pobol'še* 'he chose a somewhat bigger apple'.

(22b) **Na stole ležalo jabloko pobol'še* 'on the table there was a somewhat bigger apple'.[9]

The ideas of incitement, intention and choice are surprisingly often supported by the lexical context.

Often, it is the imperative that conveys the meaning of desire ('I want you to select among the alternatives available the one which...'), for example:

(23) *Voz'mite semgi, a ešče lučše lososiny, nu, tam, vetchiny, kolbasy, syru, kakih-nibud' konservov* ***podorože*** ≈ 'buy some salmon, preferably the best variety, ham, sausage, cheese, some preserves of the **more expensive** kind' (Ilya Ilf and Evgeny Petrov, The Wide Scope).

Sometimes, the attenuative is the only source of the meaning of selection. For example, sentence (24)

(24) *Ona nadela plat'e pojarče, čtoby vygljadet' molože* 'she put on a rather bright dress in order to look younger'

obviously conveys the idea that a dress was chosen from the brightest ones.

References

Ju.D. Apresjan, I.M. Boguslavskij, L.L. Iomdin, A.V. Lazurskij, V.Z. Sannikov, L.L. Cinman. 1992. *Lingvističeskij processor dlja složnyx informacionnyx sistem*. M.: Nauka.

Juri Apresjan, Igor Boguslavsky, Leonid Iomdin, Alexander Lazurski, Vladimir Sannikov, Victor Sizov, Leonid Tsinman. 2003. *ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT*. In: *MTT 2003, First International Conference on Meaning – Text Theory (June 16-18 2003)*. Paris: École Normale Supérieure, 2003:279-288.

I. M. Boguslavskij. 1985. *Issledovanija po sintaksičeskoj semantike*. M.: Nauka.

I.M. Boguslavskij, N.V. Grigor'ev, S.A. Grigor'eva, L.L. Iomdin, L.G. Krejdlin, N.E. Frid. 2002. *Razrabotka sintaksičeski razmečennogo korpusa russkogo jazyka*. In: *Doklady naučnoj konferencii «Korpusnaja lingvistika i lingvističeskie bazy dannyh»*. SPb: izd-vo Sankt-Peterburgskogo universiteta:40-50.

L.L. Iomdin. 2007. *Russkie konstrukcii malogo sintaksisa, obrazovannye voprositel'nymi mestoimenijami*. In: *Mir russkogo slova i russkoe slovo v mire. Materialy XI Kongressa Meždunarodnoj associacii prepodavatelej russkogo jazyka i literatury*. Heron Press, Sofia: 117-126.

L.L. Iomdin, I.A.Mel'čuk, N.V.Pertsov. 1975. *Fragment modeli russkogo poverxnostnogo sintaksisa. I. Predikativnye sintagmy*. In: *Naučno-tehničeskaja informacija*. Serija 2. # 7:30-43.

G.I. Kustova. 2002. *O semantike komparativa (absoljutivnoe upotreblenie)*. In: *Rusistika na poroge XXI veka: problemy i perspektivy. Materialy meždunarodnoj konferencii*. M., IRJa RAN: 113-120.

I.A. Mel'čuk. 1998. *Kurs obščej morfologii*. T II, Čast' 2. M.: Jazyki russkoj kul'tury, Vena: Wiener Slawistischer Almanach.

[9] It goes without saying that this sentence is absolutely correct with respect to the comparative interpretation of the attenuative.

E.V. Padučeva. 1974. O semantike sintaksisa. M., «Nauka».

Russkaja grammatika. 1980. T. 1-2. M., Nauka.

V.V. Vinogradov. 1972. *Russkij jazyk. Grammatičeskoe učenie o slove*. Izd. 2-e. M., Vysšaja škola.

Synchronous Parsing of Syntactic and Semantic Structures

Bernd Bohnet
International Computer Science Institute
1947 Center Street
Berkeley 94704, California
bohnet@icsi.Berkeley.edu

Abstract

We describe in this paper an approach for synchronous parsing of syntactic and semantic dependency structures that combines recent advances in the area to get a very high accuracy as well at the same time very good parsing times. The time for parsing, the time for training and the values of the memory footprint are to our knowledge the best results reported while the parsing accuracy are as high as the highest results reported in the 2008 shared task. The corpora used in the shared task are still different to the dependency structures of the Meaning-Text Theory. Therefore, we outline the adaption of the approach to the dependency structures of the Meaning-Text Theory.

1 Introduction

Recently, dependency parsing made large advances. Reasons for this are influential work of some researches and shared tasks for dependency parsing in the years 2006, 2007 (cf. (Buchholz and Marsi, 2006), (Nivre et al., 2007)) and the shared tasks for joint parsing of dependency and semantic structures in the years 2008 and an upcoming one in 2009 (cf. (Surdeanu et al., 2008)). There are two main approaches to dependency parsing: Maximum Spanning Tree (MST) based dependency parsing (Eisner, 1996; McDonald and Pereira, 2006) and transition based dependency parsing, cf. (Nivre and Nilsson, 2004). In this paper, we use the first approach since we could better improve the parsing speed and the MST based dependency parsing approach has a slightly better accuracy. To our knowlege there are only a few MTT parser available and even less attempts have been made to parse dependency trees of different representation levels with statistical trained parser. One of the exceptions is a transition based parser, which was train on the Russian SYNTAGRUS tree-bank, cf. (Nivre et al., 2008).

2 Parsing Algorithm

We adopted the second order MST parsing algorithm as outlined in Eisner (1996). This algorithm has a higher accuracy compared to the first order parsing algorithm since it considers also siblings and grandchildren of a node. Algorithm 1 shows the first order algorithm which is the basis for the second order parsing algorithm. Eisner (1996) first order approach can compute a projective dependency tree within cubic time (O(n^3)). Therefore, the algorithm shown in Algorithm 1 has at most three nested loops.

Both algorithms are bottom up parsing algorithms based on dynamic programming similar to the CKY chart parsing algorithm. The score for a dependency tree is the sum of all edge scores. The following equation describes this formally.

$$score(S,t) = \sum_{\forall(i,j)\in E} score(i,j)$$

The score of the sentence S and a tree t over S is defined as the sum of all edge scores where the words of S are $w_0...w_1$. The tree consists of set of nodes N and set of edges $E = \langle N \times N \rangle$. The word indices $(0..n)$ are the elements of the node set N. The expression $(i,j) \in E$ denotes an edge which is going from the node i to the node j.

The parsing Algorithm 1 searches the tree with the best score bottom up considering only the scores of single edges. Therefore, it is call first order dependency parsing algorithm. The score function in the algorithm scores always a subtree and the algorithm search for the best combination of smaller trees in order to build larger trees which maximize the overall score of the tree.

Algorithm 1: First Order Parsing Algorithm

```
// S ← w_0...w_n is the sentence with the words w_i
// l_S = n + 1 is the sentence length
// N = 0...n the indices of the words
// E = N × N set of edges
// C = (C_left, C_right, E) the chart element where C_left and C_right is a pointer to another chart element
// D = {0, 1} D represents the direction of the edge
// C_O = N × N × D × C the chart with the open sub-trees
// C_C = N × N × D × C the chart with the closed sub-trees
for j ← 1 to l_S do
  for s ← 0 to l_S do
    t ← s + j
    if t > l_S then break
    for r = s to t-1 do
      if (score(C_O[s][t][1])<score(C_C[s][r][1]) + score(C_C[r + 1][t][0])) then
        C_O[s][t][1] ← (C_C[s][r][1], C_C[r + 1][t][0]),[s→t])
      if (score(C_O[s][t][0])<score(C_C[s][r][1]) + score(C_C[r + 1][t][0])) then
        C_O[s][t][0] ← (C_C[s][r][1], C_C[r + 1][t][0]),[s←t])
    end for
    for r = s+1 to t do
      if (score(C_C[s][t][1])<score(O[s][r][1]) + score(C[r][t][1])) then
        C_C[s][t][1] ← (C_C[s][r][1], C_C[r][t][1]),[s→t])
    end for
    for r = s to t-1 do
      if (score(C_C[s][t][0])<score(C_C[s][r][0]) + score(C_O[r][t][0])) then
        C_C[s][t][0] ← (C_C[s][r][1], C_O[r][t][1]),[s←t])
    end for
  end for
end for
```

We compute the edge score ($score(i,j)$) as the scalar product of a feature vector representation of each edge $\overrightarrow{f_S}(i,j)$ with a weight vector $\overrightarrow{w}$ where i,j are the indices of the words in a sentence. The feature vector f_S might take into account not only the words with indices i and j but also additional values such as the words before and after the words w_i and w_j. The following equation shows the score function.

$$score(i,j) = \overrightarrow{f_S}(i,j) * \overrightarrow{w}$$

Most of the features are built out of the properties from the words such as part-of-speech, morphologic features, lemmas, and forms. For instance, some of the features for the words 2 (bought) and 5 (computer) of the sentence below would be {`VB+N, VB+N+Distance_3, buy+N, VB+computer, buy+computer`}. The features are frequently encoded as strings and mapped to a number. The number becomes the index of the feature in the feature vector and weight vector. Therefore, a feature vector looks like the following expression $f_S(2,5) = \{0,0,0,1,0,0,..,0,1,..,1,..1,..\}$ [1] and the weight vector looks

[1] A implementation of such a sparse vector has to store the values more efficient.

similar, e.g. $w = \{0.1258, -0.2554, 0, 0.333, -0.0125, ...\}$.

$$\text{He}_1 \text{ bought}_2 \text{ a}_3 \text{ new}_4 \text{ computer}_5 \text{ which}_6 \text{ was}_7 \text{ very}_8 \text{ expensive}_9 \text{ .}_{10}$$

In order to compute the weight vector, we use a support vector machine which has proven to be efficient for dependency parsing. The support vector machine implements online large margin multi-class learning, cf. (Crammer et al., 2003; McDonald et al., 2005). We provide more details in Section 3.

2.1 Second-Order Dependency Parsing

The first order dependency parsing algorithm takes only into account the parent and one dependent. While the second order algorithm uses information of the composition of the subtrees namely the edges of grandchildren and siblings. This improves the parsing accuracy since for instance an edge to a preposition has to be followed mostly by an edge to a noun to complete the prepositional part and also coordination consist always of more than one edge.

2.2 Labeled Dependency Parsing

Algorithm 1 builds an unlabeled dependency tree. However, all new dependency tree banks have trees with labeled edges. The following two approaches are common to solve this problem: The first approach uses an additional algorithm to label the edges in a post-processing step. The second approaches extends the parsing algorithm and integrates the labeling algorithm into the parsing algorithm.

McDonald et al. (2006) use an additional algorithm. Their two stage model has a good computational complexity since the labeling algorithm contributes again only a cubic time complexity to the algorithm (O(n^3)) and keeps therefore the joint algorithm still cubic. This solution computes the edge labels for each possible edge separate from the unlabeled dependency tree. The algorithm has three loops. The first two loops iterate over the words of the sentence and they build a matrix which refer to all possible edges i, j. The third loop iterates over all possible labels and selects the highest scored label due to the score function $score(w_i, label) + score(w_j, label)$ and inserts the highest scored label into the matrix. The scores are also used in the parsing algorithms and added to the edge scores which improves the overall parsing results as well. In the first order parsing scenario, this procedure is sufficient since no combination of edges are considered by the parsing algorithm. However, in the second order parsing scenario where more than one edge are considered by the parsing algorithm, combinations of two edges might be more accurate than two single edges with the highest score.

Carreras (2007) as well as Johansson and Nugues (2008) combine edge labeling with the second order parsing algorithm. This adds an additional loop over the edge labels to the parsing algorithm. The complexity is therefore O(n^4). This increases in our experiments the parsing accuracy from about 0.86 labeled accuracy score to 0.88 what is a relative high improvement. We got the first value for our partly reimplementation of the McDonald and Pereira (2006) parser and the second value for the parsing algorithm that includes labels. For our experiments, we used the English dependency tree bank as provided in the CoNLL shared task 2008.

2.3 Non-Projective Dependency Parsing

The dependency parsers developed in the last few years use two different techniques for non-projective dependency parsing. The most common technique uses tree rewriting and was invented by Kahane et al. (1998). This technique was taken up again by Nivre and Nilsson (2005) in Nivre's transition based dependency parser which performed second best in the 2006 shared task which has therefore become well known.

By using pseudo-projective dependency parsing, the training data for the parser is first projectivized by applying a minimal number of lifting operations to the non-projective edges and encoding information about

these lifts in edge labels. After this operations, the trees are projective and therefore a projective dependency parser can be applied. During the train, the parser learns also to built trees with the lifted edges and so indirect to built non-projective dependency trees since after the projective dependency parsing the inverse operations to the lifting are performed and by this operation the edges are moved downwards the tree and non-projective trees are built.

McDonald and Pereira (2006) developed a technique to rearrange edges in the tree in a postprocessing step after the projective parsing has taken place. They call the algorithm approximation non-projective dependency parsing. It searches first the highest scoring projective parse tree and then it rearranges edges in the tree, in each step one, until the rearrangements does not anymore increase the score for the tree. This technique is computationally expensive for trees with a large number of non-projective edges since it considers to reattach all edges to any other node until no higher scoring trees can be found. Their argument for the algorithm is that most edges in a tree even in language with lot of non-projective sentences, the portion of non-projective edges are still small and therefore by starting with the highest scoring projective tree, typically the highest scoring non-projective tree is only a small number of transformations away from.

Threshold	Labeled Accuracy Score (LAS)	Unlabeled Accuracy Score (UAS)
projective	0.86711	0.90350
0.0	0.86653	0.90127
0.8	0.86727	0.90416
1.0	0.86852	0.90619
1.1	0.86880	0.90530
1.2	0.86790	0.90421

Table 1: Accuracy Scores of different thresholds for the approximation non-projective dependency parsing.

We found out in our experiments with the non-projective approximation algorithm that with a threshold higher then about 0.7, the parsing accuracy even for English slightly improves. With a threshold of 1.1, we got the best improvement. The results of this experiment are summarized in Table 1. Since we opt for a high labeled accuracy score, we selected a threshold of 1.1 also if the unlabeled score for 1.0 is higher.

3 Parsing Framework

One of the main goals of the paper is to show how such a parser can be implemented fast without loosing accuracy. This is very important for the applications which use a parser. Parsing is in most cases only one component of a Natural Language Processing application such as summarization, dialog system or machine translation. Many applications have a tight time schedule and a parser can not take more than a second or even only some milliseconds to parse a sentence. The same holds true for the memory footprint since the parser has to share the memory with other components of a system or it has to run on a small device which might provide only a very limited amount of memory. During the development of a parser, it is very important as well that it does not take too long to train the parser since otherwise experiments take too long and it becomes impossible to improve the parser in a give time.

3.1 Online Learning

As learning technique, we use Margin Infused Relaxed Algorithm (MIRA) as developed by Crammer et al. (2003) and applied to dependency parsing by McDonald et al. (2005). The Algorithm in Figure 2 processes one training instance on each iteration, and updates the parameters due to the currently processed instance.

The inner loop iterates over all sentences x of the training set while the outer loop repeats the train i times. The algorithm returns an averaged weight vector and uses an auxiliary weight vector v that accumulates the values of w after each iteration. At the end, the algorithm computes the average of all weight vectors by dividing it by the number of training iterations and sentences. This helps to avoid overfitting, cf. (Collins, 2002).

Algorithm 2: MIRA

$\tau = \{S_x, T_x\}_{x=1}^{X}$ // The set of training data consists of sentences and the corresponding dependency trees
$\overrightarrow{w}^{(0)} = 0$, $\overrightarrow{v} = 0$
for n = 1 **to** N
  for x = 1 **to** X
    w^{i+1} = update w^i according to instance (S_x, T_x)
    $v = v + w^{i+1}$
    $i = i + 1$
  end for
end for
$w = v/(N * X)$

The update function computes the update to the weight vector w^i during the training so that wrong classified edges of the training instances are possibly classified correct. This is computed by increasing the weight for the correct features and decreasing the weight for wrong features of the vectors for the tree of the training set $\overrightarrow{f_{T_x}} * w^i$ and the vector for the predicted dependency tree $\overrightarrow{f_{T'_x}} * w^i$.

The update function tries to keep the change to the parameter vector w^i as small as possible for correctly classifying the current instance with a difference at least as large as the loss of the incorrect classifications. The update of the algorithm in Figure 2 can be formalized by following update function.

$$\min ||w(i+1) - w(i)||$$
$$s.t.\ score(S_x, T_x) - score(S_x, T'_y) \geq L(T_x, T'x))$$

3.2 Selected Parsing Features

Table 2 and 3 give an overview of the selected features for the reimplementation of McDonald and Pereira (2006) (System A) and the extended version with the integrated edge labels (System B), cf. Johansson and Nugues (2008). Both parser versions shared the same code and training algorithm except from two parts: the parsing algorithm itself which are only about 100 lines of code and the code which extract the features. The reason for this is that the parsing accuracy of both algorithms are sensitive to the selected features. For the parsing and training speed, most important is a fast feature extraction beside of a fast parsing algorithm.

3.3 Implementation Aspects

In this subsection, we provide some implementation details that concern all the speed of the parser and distinguish this implementation from others. The learning architecture determines the architecture of the parser.

The training has three passes. The goal of the first two passes is to collect the set of possible features for all elements of the training set. In the first pass, the feature extractor collects all attributes that the features can contain since our goal is to determine the minimal description length for each attribute. The reason for this is to save memory and computational time during the feature creation. For instance, when the feature extractor builds features due to the pattern `label, h-pos, d-form`[2] then in the first pass the attributes edge labels, part-of-speech tags (pos) and word forms are included into collection procedure. For each category (labels, pos, etc.), the extractor builds a mapping to a number which is continuos from 1 to the count of elements without duplicates. The following equation shows this formally.

We enumerate in the same way the feature patterns and obtain the function $f_{feature-type}(value)$, e.g. $f_{feature-type}$(label,h-pos,d-form)=7. Now, we can calculate the minimal description length in bits for each

[2]We combine all the elements (label, pos, form) of this feature pattern to a single feature (label+pos+form).

Standard Features				**Linear Features**	
Feature	System	Feature	System	Feature	System
h-form	A,B	h-form, d-pos	A,B	h-pos, d-pos, h-pos + 1	A,B
h-pos	A,B	h-pos, d-form	A,B	h-pos, d-pos, h-pos - 1	A,B
d-form	A,B	h-form, d-form	A,B	h-pos, d-pos, d-pos + 1	A,B
d-pos	A,B	h-pos, d-pos	A,B	h-pos, d-pos, d-pos - 1	A,B
h-form,h-pos	A,B	h-form, d-form, h-pos	A,B	h-pos, d-pos, h-pos - 1, d-pos - 1	A,B
d-form,d-pos	A,B	h-form, d-form, d-pos	A,B	h-pos, d-pos, h-pos - 1, d-pos + 1	A,B
		h-pos, d-pos, h-form	A,B	h-pos, d-pos, h-pos + 1, d-pos - 1	A,B
		h-pos, d-pos, d-form	A,B	h-pos, d-pos, h-pos + 1, d-pos + 1	A,B
		h-pos, d-pos, h-form, d-form	A,B	h-pos - 1, d-pos, d-pos + 1	A
				h-pos + 1, d-pos, d-pos - 1	A
				h-pos - 1, h-pos, d-pos + 1	A
				h-pos + 1, h-pos, d-pos - 1	A

Grandchild Features		**Sibling Features**	
Feature	System	Feature	System
h-form, d-pos, g-pos	A	d-form, s-form $\oplus$ dir(d,s) $\oplus$dist(d,s)	A
h-form, d-pos, g-pos, dir(h,d)	A	d-pos, s-form $\oplus$ dir(d,s) $\oplus$dist(d,s)	A
		d-pos, s-form $\oplus$ dir(d,s)+ $\oplus$ dist(d,s)	A
		d-pos, s-pos $\oplus$dir(d,s)$\oplus$dist(d,s)	A
h-pos, d-pos, g-pos, dir(h,d), dir(d,g)	B	h-pos, d-pos, s-pos, dir(h,d), dir(d,s) $\oplus$dist(h,s)	B
h-form, g-form, dir(h,d), dir(d,g)	B	h-form, s-form, dir(h,d), dir(d,s)$\oplus$dist(h,s)	B
d-form, g-form, dir(h,d), dir(d,g)	B	d-form, s-form, dir(h,d), dir(d,s) $\oplus$dist(h,s)	B
h-pos, g-form, dir(h,d), dir(d,g)	B	h-pos, s-form, dir(h,d), dir(d,s)$\oplus$dist(h,s)	B
d-pos, g-form, dir(h,d), dir(d,g)	B	d-pos, s-form, dir(h,d), dir(d,s)$\oplus$dist(h,s)	B
h-form, g-pos, dir(h,d), dir(d,g)	B	h-form, s-pos, dir(h,d), dir(d,s)$\oplus$dist(h,s)	B
d-form, g-pos, dir(h,d), dir(d,g)	B	d-form, s-pos, dir(h,d), dir(d,s)$\oplus$dist(h,s)	B

Table 2: Selected Features. h stands for head, d for dependent, g for grandchild, and s for sibling. System A builds additional features by adding the **dir**ection and a feature that has additional the **dist**ance plus the direction. The direction is left if the dependent is left of the head otherwise right. The distance is the number of words between the head and the dependent, if $\leq$5, 6 if $>$5 and 11 if $>$10. In some cases, we could also improve system B by adding this features as well. In this cases, we list this explicit. System B has always the edge label included in the features which is not indicated in order to make to compare easier. $\oplus$ means that an additional feature is build with the previous part plus the next part.

$f_{attribute}(value) \rightarrow N$, e.g. let be $f_{labels}(value) \rightarrow \{$(punc,0),(sbj,1),(obj,2),(mod,3) ..$\}$ then $f_{label}(sbj) = 1$ [3]

of the attributes with the following equation:

$$bits(attribute) = ceil(log_2(max(N_{attribute})))$$

In the second pass, the extractor builds the features for all training examples which occur in the train set but not for all combination, i.e., the extractor builds feature for all in the training set contained edges. In other words, only for the positive examples and not for the negative cases that do not occur. However, these features of the *wrong edges* could improve the parser accuracy since the parser considers also this edges during the creation of the parse tree. This would lead to a much larger number of features. Therefore, most of the implementation do not consider these features.

We create the features with a function that maps iteratively the attributes of a feature to a number represented with 64 bit and then enumerates and maps these numbers to 32 bit numbers to save even more memory. This is computed by the equation $l(value, start, value_{previous}) = last_{previous} + (value << start)$. The expression `number<<n` means shift the binary representation of the number by n-bits to the left. For instance, let *mod,N* be the set of attribute of the feature pattern `label, h-pos`, $bits$(labels)=7, $bits$(pos)=6, $bits$(feature-type)=6, f_{labels}(mod)= 3 (11b), f_{pos}(N)=6(110b), and $f_{feature-type}(label + h - pos)$=8

Label Features			
Feature	System	Feature	System
label, pos $\oplus$ child $\oplus$ dir(h,d)	A	label, pos, pos + 1 $\oplus$ child $\oplus$ dir(h,d)	A
label, pos, pos-1, pos-2 $\oplus$ child $\oplus$ dir(h,d)	A	label, pos, pos -1, pos-2, pos+1 $\oplus$ child $\oplus$ dir(h,d)	A
label, pos, pos-2, $\oplus$ child $\oplus$ dir(h,d)	A	label, pos, pos +2, $\oplus$ child $\oplus$ dir(h,d)	A
label, pos, pos-1 $\oplus$ child $\oplus$ dir(h,d)	A	label, pos, pos+1 $\oplus$ child $\oplus$ dir(h,d)	A
label, pos, pos+1, pos-1 $\oplus$ child $\oplus$ dir(h,d)	A	label, form $\oplus$ child $\oplus$ dir(h,d)	A

Table 3: System A uses a boolean flag **child** to indicate that it is the head or dependent and adds these feature once for the head and once for the dependent. $\oplus$ means that an additional feature is build with the previous part plus the next part.

(1000b). Then the value for the feature type is computed by $l(8,0,0) = 8$ (1000b) and start=$bits$(feature-type)=6; l(3,6,8)= 1000b + (11b<<6) = 1000b + 11000000b = 11001000b =200 and start=6+$bits$(label)=12; $l(12,6,200)$= 11001000b + (11b<<12)= 11001000b + 11000000000000 = 11000011001000b.

The following list shows an overview of the most important implementation details that improve the speed:

1. We use as feature vector within the support vector machine only a list of the features without any additional (double floating point) value.

2. We store the feature vectors for $f(label, w_i, w_j)$, $f(label, w_i, w_j, w_g)$, $f(label, w_i, w_j, w_s)$ etc. in a **compressed** file. The reason for storing vectors in a file is that it is faster to compute the values once and then to load them in each of the training iterations (6-10 times).

3. We zip the file with the option for fast compression and decompression. The reason for this is that otherwise the IO to the hard disk drive becomes the bottleneck.

4. After the training, we store only the parameters of the support vector machine which are not zero.

	McDonald and Pereira (2006)	System A	Johansson and Nugues (2008)	System B
Type	2nd order	2nd order	2nd order integrated labels	2nd order integrated labels
training time	70 hours	4 hours	60 hours	14 hours
memory usage	7 GB	1.5 GB	not reported	3 GB
parsing time	2 seconds	0.05 seconds	1.49 seconds	0.6 seconds
memory usage	1.5 GB	700 MB	not reported	1 GB
LAS	0.86	0.87	0.88	0.88

Table 4: Performance Comparison

Table 4 gives an overview of different parsing systems and their performance and memory usage. We use the training and development set of the 2008 shared task and for the speed comparison a 2.8 Ghz Mac Pro and the values reported in Johansson and Nugues (2008) based on 3.2 Ghz Mac Pro.

4 Semantic Role Labeling

Semantic Role Labeling (SRL) as well as Dependency Parsing has been topics of CoNLL shared tasks, cf. (Carreras and Màrques, 2004; Carreras and Màrquez, 2005). The first two shared task 2004 and 2005 used phrase structures trees as input to the semantic role labeler while the last shard task (2008) and the upcoming CoNNL shared task (2009) uses dependency trees. We use a pipeline architecture for semantic role labeling. The components of the pipeline are predicate identification (PI), argument identification (AI), argument classification (AC), and word sense disambiguation (WSD). For training and testing, we use the English

corpus of the shared task 2008. The corpus is in addition to dependency trees annotated with predicates and semantic roles of the NomBank and PropBank, cf. (Meyers et al., 2004; **?**).

Algorithm 3:Attribute Identification

// $S_x \leftarrow w_0...w_n$ is the sentence x with the words w_i
// $P_x \leftarrow p_0...p_m$ is the set of predicates p_j of sentence x
A_j^x is the set of arguments of predicate j in sentence x.
for all $p_j \in P_x$
  for all $w_i \in S_x$
    if score$(p_j, w_i) \geq 0$ **then** $A_j^x \leftarrow A_j^x \cup \{i\}$

In order to identify the predicates, we look up the lemmas in the PropBank and NomBank. For all other components, we use the same learning technique and architecture as for the dependency parser. We use the same technique because we want to be able to use the scores of the components to rerank other results and the used support vector machine allows a very large number of features that standard decision trees, neural nets, etc. can not handel.

Algorithm 3 identifies the arguments of each predicate. Its two loops iterate over the predicates and over the words of a sentence in the case that the score function is large or equal to zero the argument is added to the set of arguments of the predicate in question.

The argument classification algorithm labels each argument identified in the previous step with a semantic role label. The argument classification algorithm selects with a beam search algorithm the combination of arguments with the highest score. The algorithm allows only one core argument of the same type such (A0 to A5).

The last component of our pipeline is the word sense disambiguation. We put this against the intuition at the end of our pipeline since experiments showed that other components could not profit from disambiguated word senses but on the other hand the word sense disambiguation could profit from the argument identification and argument classification. In order to disambiguate, we iterate over the words in the corpus that have more than one sense and take the sense with the highest score. Due to space restrictions, we can not list all the features that we used in our systems. A lot of good combinations can be found in Che et al. (2008).

The accuracy and scores of our system are only a bit lower than the best reported results (80.4) on the development data of the 2008 shared task, cf. Johansson and Nugues (2008). The Attribute Identification component has a accuracy of 91.6 and the attribute classification applied on the output of the AI a F1 score of 77.5 Since we consider at this stage that all word sense have the first sense 01, the Word Sense Disambiguation can improve the results to 79.2. The average time to execute the SRL pipeline on a sentence is less than 0.15 seconds.

5 Application to the Meaning-Text Theory

The above technique could be directly applied to a MTT corpus. The dependency trees converted from the phrase structure annotation of the Penn Treebank have become much more similar to the surface syntactic trees for instance the coordinations are no longer flat and attached to the conjunction. In a lot of cases, only the edge labels are different. Therefore, the described techniques could be applied to a corpus annotated with MTT surface syntactic dependency trees.

The mapping of surface syntactic dependency trees to deep syntactic dependency trees can be addressed with similar techniques. In this step the main task is to leave out the function words, to introduce lexical functions and to label the edges with deep syntactic dependency labels.

The semantic graphs of the MTT are mostly comparable to the PropBank annotation. The exceptions are the communicative structures which is missing and predicates which form lexical functions are represent different on the semantic stratum.

We hope that MTT Corpora annotated with dependency representation of all levels become available: surface syntactic structures, deep syntactic structures and semantic representations including the communicative structure (Mel'čuk, 2001). We are sure that this would be one of the most valuable linguistic resource. One of the most recent initiative towards this direction is a corpus for Spanish, cf. (Mille et al., 2009).

6 Conclusion

In this paper, we have described an algorithm for synchronous parsing of syntactic and semantic structures. Our implementation has scores that are comparable good as the best so far reported results. Moreover, the implementations and techniques introduced in this paper provide a much better parsing and training times. Also the memory footprint are lower so that the parsers can be trained on standard computer and used on devices which have less memory.

We integrated the synchronous parser into the Meaning-Text Development Environment (Mate) (Bohnet et al., 2000) which can be trained now on MTT corpora so that it is possible to obtain surface syntactic dependency trees and the semantic actants with the above technique when trained on a corpus annotated with MTT structures. This can help also to set up corpora in a boots trap approach.

References

Bohnet, B., A. Langjahr, and L. Wanner. 2000. A Development Environment for an MTT-Based Sentence Generator. In *Proceedings of the First International Natural Language Generation Conference*.

Buchholz, S. and E. Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *In Proc. of CoNLL*, pages 149–164.

Carreras, X. and L. Màrques. 2004. Introduction to the conll-2004 shared task: Semantic role labeling. In *Proceedings of CoNLL-2004*, pages 89–97. Boston, MA, USA.

Carreras, X. and L. Màrquez. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 152–164, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Carreras, Xavier. 2007. Experiments with a Higher-Order Projective Dependency Parser. In *Proceedings of the EMNLP-CoNLL 2007 Shared Task*.

Che, Wanxiang, Zhenghua Li, Yuxuan Hu, Yongqiang Li, Bing Qin, Ting Liu, and Sheng Li. 2008. A Cascaded Syntactic and Semantic Dependency Parsing System. In *CoNLL 2008: Twelfth Conference on Computational Natural Language Learning*, pages 238–242, Manchester, England. Coling.

Collins, M. 2002. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In *EMNLP*.

Crammer, K., O. Dekel, S. Shalev-Shwartz, and Y. Singer. 2003. Online Passive-Aggressive Algorithms. In *Sixteenth Annual Conference on Neural Information Processing Systems (NIPS)*.

Eisner, J. 1996. Three New Probabilistic Models for Dependency Parsing: An Exploration. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, pages 340–345, Copenhaen.

Johansson, R. and P. Nugues. 2008. Dependency-based syntactic–semantic analysis with PropBank and NomBank. In *Proceedings of the Shared Task Session of CoNLL-2008*, Manchester, UK.

Kahane, S., A. Nasr, and O. Rambow. 1998. Pseudo-projectivity: A polynomially parsable non-projective dependency grammar. In *COLING-ACL*, pages 646–652.

McDonald, R. and F. Pereira. 2006. Online Learning of Approximate Dependency Parsing Algorithms. In *In Proc. of EACL*, pages 81–88.

McDonald, R., K. Crammer, and F. Pereira. 2005. Online Large-margin Training of Dependency Parsers. In *Proc. ACL*, pages 91–98.

McDonald, R., K. Lerman, F. Pereiraand, K. Crammer, and F. Pereira. 2006. Multilingual Dependency Parsing with a Two-Stage Discriminative Parser. In *Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 91–98.

Mel'čuk, I.A. 2001. *Communicative Organization in Natural Language : The Semantic-Communicative Structure of Sentences*. John Benjamins Publishing, Philadelphia.

Meyers, A., R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004. The nombank project: An interim report. In Meyers, A., editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 24–31, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Mille, S., V. Vidal, A. Burga, and L. Wanner. 2009. Creating an MTT Tree Bank of Spanish. In *Proceedings ot the Fourth International Conference on Meaning-Text Theory*, Montréal.

Nivre, J., Hall J. and J. Nilsson. 2004. Memory-Based Dependency Parsing. pages 49–56, Boston, Massachusetts.

Nivre, J. and J. Nilsson. 2005. Pseudo-Projective Dependency Parsing. In *In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 99–106.

Nivre, J., J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel, and D. Yuret. 2007. The conll 2007 shared task on dependency parsing. In *Proc. of the CoNLL 2007 Shared Task. Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, June.

Nivre, Joakim, Igor M. Boguslavsky, and Leonid L. Iomdin. 2008. Parsing the SYNTAGRUS Treebank of Russian. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 641–648, Manchester.

Surdeanu, M., R. Johansson, A. Meyers, L. Màrquez, and J. Nivre. 2008. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the 12th Conference on Computational Natural Language Learning (CoNLL-2008)*.

PROPRIETES COMBINATOIRES ET POLYSEMIES DE BASES VERBO-NOMINALES EN WOLOF : QUELLE CORRELATION ?

Olivier Bondéelle
Laboratoire Modèles Dynamiques Corpus - Université Paris 10
Leiden University Centre for Linguistics - Université de Leiden
10 rue de Panama 75018 Paris France
olivbond@yahoo.fr

Resume

Cet article examine la possibilité d'extraire des propriétés combinatoires de bases verbo-nominales, qui peuvent contribuer à décrire la polysémie de ces bases.
L'article comporte trois parties. Dans la première, je décris la compatibilité de verbes polysémiques des émotions du wolof avec l'aspect inaccompli à valeur d'inchoatif. La seconde partie est consacrée à la compatibilité des lexèmes nominaux avec des marques de pluriel et avec des lexèmes qui dénotent des quantités. L'analogie entre nom massif et verbe d'état est la perspective de cet article. La dernière partie pose la question de la formalisation de cette analogie sémantique entre les verbes et les noms dans un modèle Sens-Texte.

Abstract

This paper explores the possibility to extract some combinatorial properties of verb-noun roots, which can contribute to describe the polysemy of these roots.
This paper contains three parts. In the first one, I describe the compatibility of polysemic verbs of emotion in Wolof with the inchoative aspect. In the second one, I describe the compatibility of nouns of emotion with the plural forms and either with lexemes which refer to quantities. The analogy between mass nouns and state verbs is the perspective of this paper. The last part asks how to formalize this semantic analogy in a Meaning-Text model.

Introduction

En wolof, on appelle bases verbo-nominales les lexèmes soumis à une conversion[1] et qui ne sont pas déjà dérivés morphologiquement (Nouguier-Voisin, 2002 : 19) : « En effet, une grande part des unités lexicales du wolof sont mieux identifiées comme des bases verbo-nominales, c'est-à-dire des unités lexicales qui, sans avoir à subir de dérivation, sont également aptes à fonctionner comme base nominale ou comme base verbale ». On en trouve une vingtaine dans le champ sémantique des émotions parmi lesquelles[2] : JAAXLE '(être) inquiet, étonné, embarrassé, être dans une situation embarrassante ; (G) inquiétude' ou MER 'être en colère, se mettre en colère ; (M) colère' ou encore RAGAL 'avoir peur,

[1] Mel'čuk (1996 : 133) a caractérisé la conversion catégorielle : « ce qui change est la CLASSE SYNTAXIQUE MAJEURE (lettres capitales de l'auteur) du radical de départ – sa partie du discours. Cela entraine simultanément le changement de l'ensemble des constructions syntaxiques qui admettent les mots-formes avec ce radical ainsi que (s'il y a lieu) celui de l'ensemble des catégories flexionnelles pertinentes à ce dernier. »

[2] La lettre en majuscules et entre parenthèses indique la classe nominale à laquelle appartient le nom. Le wolof est une langue à classe nominale. Ainsi, RAGAL G- 'la peur' et RAGAL B- 'le peureux'.
Les dictionnaires wolof-français que j'utilise (Fal et al, 1990 et Diouf, 2001) ne donnent pas de définition des entrées lexicales. C'est pourquoi je ne hiérarchise pas ici les lexies des vocables. Il faudrait le faire en toute rigueur comme dans une modélisation Sens-Texte aboutie. Ici, la virgule sépare juste les sens différents.

prendre peur ; (G) peur ; (B) peureux' et SEDD ' être froid, faire froid; (B) froid, froideur, rhumatismes; être calme; (G) calme; être satisfait, être frigide 'et enfin TIIS 'être triste, faire de la peine; (W) malheur, souci'. Ces bases présentent une polysémie régulière du type *action* : *nom d'action*[3] ou *action : nom d'actant* (Apresjan, 1992 : 216-259) due à la conversion. Certains lexèmes verbaux comme MER, RAGAL présentent aussi une polysémie du type 'être dans un état émotionnel' : 'commencer à être dans un état émotionnel' où le sens statif alterne avec le sens inchoatif (Apresjan, 1992 : 70). Le lexème nominal semble à priori ne sélectionner que le sens statif du lexème verbal. Pour les conversions catégorielles avec des verbes d'émotion qui sont polysémiques (sens statif et sens inchoatif), Apresjan (1992 : 182) a proposé que lorsque le verbe a le sens statif, le nom est du type S_0 alors qu'il est du type S_{RES} lorsque le verbe a le sens inchoatif. Ainsi, S_0 (MER 'être en colère') = MER M- 'colère' et S_{res} (MER 'se mettre en colère') = MER M- 'colère'.

Peut-on extraire des propriétés combinatoires de ces verbes qui aideraient à décrire ces polysémies ? Les noms possèdent-ils des propriétés similaires ? Qu'en est-il alors des propriétés des bases verbo-nominales ? Commençons par présenter la combinatoire des verbes d'émotion au niveau morphosyntaxique. Pour cela, il est nécessaire de caractériser les catégories flexionnelles de ces verbes.

1 Une propriete combinatoire aspectuelle des verbes

1.1 Point de vue aspectuel pour distinguer les etats des actions

En wolof, le point de vue aspectuel permet de distinguer les états des actions. Avant de considérer le cas du parfait, il faut noter que tout lexème verbal dans un énoncé est associé à un morphème grammatical flexionnel. Ce morphème est complexe puisqu'il amalgame des indices de personne, de mode et d'aspect. Le système verbal du wolof lui-même est complexe et ne peut pas être présenté ici.[4] Je me bornerai à en présenter des aspects primaires qui permettent de mieux saisir le comportement des verbes d'émotion. La forme et la position du morphème grammatical par rapport au lexème verbal permettent de préciser l'aspect et le mode du verbe. La plupart du temps, il est antéposé au lexème verbal comme en (1a-b), sauf au parfait comme en (1c). En (1a), il y a seulement un indice de personne antéposé au lexème verbal: le verbe est donc à l'aspect perfectif et au mode narratif. En (1b), le morphème est antéposé au lexème verbal mais constitué de deux indices: un indice de mode *da-* et un indice de personne *–ma*. En wolof, ce n'est que l'aspect imperfectif qui est marqué par un grammème spécifique. L'aspect perfectif est un grammème zéro. Ici, le verbe est à l'aspect perfectif et au mode de l'emphatique du verbe (EV dorénavant).[5]

(1a)	**Ma**	gis	li	nga	yor		/	(1b)	**dama**	gis	sa	gàllaaj	yi
	1SG	voir	ce	2	avoir	avec	/		**EV.1SG**	voir	POSS.2SG	gris-	PL
			que	SG	soi							gris	
	'J'ai vu ce que tu avais avec toi'						/		'J'ai vu tes gris-gris'				

[3] Les deux points représentent une proximité sémantique entre deux sens sans préjuger de la direction du changement de sens comme dans Apresjan (1992).

[4] Le livre de Robert (1991) est une présentation détaillée de ce système.

[5] **CONVENTIONS**

CL: classe nominale
DEF: défini
ECPT : emphatique du complément
EV: emphatique du verbe
INAC : inaccompli
INDEF : indéfini
LOC : locatif
MOD : modalisateur
NARR : narratif
NEG : négatif
PARF: parfait
PL : pluriel
POSS : possessif
PRES : présentatif
REL : relateur
SG : singulier

(1c) **Gis** **naa** Faal Faadel
voir **PARF.1SG** Fal Fadel
'J'ai vu Fadel Fal'

Il est important de noter comme le fait Robert (2002) que « la valeur temporelle ne suffit pas à définir les conjugaisons ». A part le présentatif dont la forme est *a ngi*, qui a obligatoirement valeur de présent (1d), les autres modes ont potentiellement valeur de passé. Il faut s'arrêter sur le parfait car la traduction varie en fonction du verbe utilisé. En wolof, on distingue des verbes selon leur comportement avec telle ou telle conjugaison. Un verbe est dit verbe d'état lorsque son emploi avec le parfait et avec l'emphatique du verbe ont valeur de présent temporel. En revanche, un verbe d'action peut être employé au présentatif et l'emploi avec le parfait a une valeur d'état résultant (l'emploi avec l'emphatique du verbe reste plus ambigu car cette conjugaison donne une importance particulière aux valeurs modales et pragmatiques). Cette partition parmi les types de verbes est cohérente avec celle qu'établit Bach (1986) parmi les évènements qu'il nomme en anglais *eventualities* puisque la partition majeure dans sa classification sépare nettement états et non-états.[6] En wolof, le parfait en (1c) donne un exemple de résultatif avec un verbe d'action, alors que son utilisation avec un verbe d'état comme en (1e-f) n'a pas forcément une valeur résultative.

(1d)[7] Baadoolo ya-**a** **ngi** leen jeexal
Pauvre PL-**MOD$_1$** **MOD$_2$** 3PL ruiner
'Les pauvres ils les ruinent'

(1e) Sonn naa / (1f) baax na
être fatigué **PARF.**1SG / être bien **PARF.**3SG
'Je suis fatigué' / 'C'est bien'

Dans ces conditions, il faut attacher une importance particulière à la marque *-y* d'inaccompli suffixé au morphème de personne, aspect et mode (je néglige ici sa valeur modale qui est pourtant à prendre en compte en toute rigueur cf. Robert, 1991 : 265-269). Pour un verbe d'action, la marque d'inaccompli donne au procès une valeur de présent progressif comme en (2a) ; ou une valeur d'habitude au narratif comme en (2b).

(2a) Dama-**y** lekk céeb / maa ngi-y lekk céeb
EV.1SG-INAC manger riz / **1SG.MOD$_1$** **MOD$_2$-INAC** manger riz
'Je **suis en train** de manger du riz'

(2b) **Ma-y** seetsi sama ëriñ bès bu nekk
1SG-INAC rendre visite POSS.1SG marabout jour chaque
'Je **rends visite** à mon marabout chaque jour'

En revanche avec les verbes d'état de qualité, la forme d'inaccompli *–y* n'est pas admissible pour exprimer présent progressif comme en (2c). Seuls l'habitude et le futur sont possibles.

---

[6] Bach utilise les termes anglais *event* et *process* pour désigner les non-états. Pour éviter les confusions avec « évènement » et « procès » en français, j'utilise « action » pour désigner les non-états.

[7] Le présentatif est un mode difficile à analyser en termes morphologiques. Il combine plusieurs marqueurs que je note **MOD$_1$** et **MOD$_2$** . Les analyses divergent sur le statut de ces marqueurs cf. Robert (1991 : 169-170). Dans certaines analyses, **MOD$_1$** serait une marque d'emphase de personne tandis que d'autres privilégient l'idée d'un indice spatial. D'autres analyses militent plutôt en faveur d'un marqueur unique de modalité qui réunit **MOD$_1$** et **MOD$_2$** . Dans un souci de simplicité, j'ai choisi de mettre un numéro en indice du modalisateur sans établir de différence de statut entre **MOD$_1$** et **MOD$_2$**.

(2c) Senegal **dafa** tàng / Senegal **dafa-y** tàng
Sénégal **EV.3SG** être chaud / Sénégal **EV.3SG-INAC** être chaud
'Au Sénégal, il fait chaud' / 'Au Sénégal, il va faire chaud ; il fait chaud (habituellement, mais pas en ce moment précis)

On peut résumer la dichotomie qui existe entre verbes d'état et verbes d'action du seul point de vue aspectuel. Le parfait donne au verbe d'action une valeur obligatoire de résultatif alors que ce n'est pas le cas avec un verbe d'état. Et la forme de l'inaccompli *–y* a une valeur de présent progressif ou d'habitude avec un verbe d'action alors que la valeur progressive de l'inaccompli *–y* n'est pas admissible avec un verbe d'état. Suivant la classification de Bach, le procès d'état statif n'admet pas d'être découpé dans le temps. Il est homogène et Bach le caractérise par son caractère continu.[8] Par contraste, l'action (selon notre appellation) est dynamique et le procès est sécable dans le temps. Le verbe d'action exprime un procès hétérogène et discret linguistiquement. En wolof, le procès exprimé par le verbe d'état de qualité n'admet pas le présent progressif (ici l'inaccompli).

Examinons maintenant comment se comportent les verbes d'émotion par rapport au parfait (donc nécessairement à l'aspect perfectif) et à l'inaccompli *–y* (aspect imperfectif) pour savoir si la langue wolof les considère comme des procès discrets ou continus.

1.2 Le comportement combinatoire des verbes d'emotion

Le comportement combinatoire des lexèmes verbaux des émotions avec le parfait et avec la marque d'inaccompli est intéressant du point de vue du type d'évènement. Il faut noter que le verbe d'émotion se comporte comme un verbe d'état puisque la valeur du parfait et de l'emphatique du verbe n'est pas nécessairement le résultatif comme en (3a-b), ce qui est une caractéristique des verbes d'état comme on l'a noté.

(3a) Góor ñi **rus** **nañu** a wone seen ragal
homme **PL** **avoir honte** **PARF.3PL** MOD montrer POSS.3PL peur
'Les hommes **ont honte** de manifester leur peur'

(3b) Omar **dafa** sedd
Omar **EV.3SG** être froid
'Omar **est calme**', litt. 'Omar il est froid'

Pourtant, la combinaison avec la forme *–y* de l'inaccompli quand elle prend la valeur de l'habitude est acceptable dans certaines conditions mais fortement contrainte (3c-d)

(3c) Su ma tëddee guddi ma geestu yem
si 1SG se coucher nuit 1SG retourner la tête pour regarder tomber sur

ci Laay **dama-y** **kontaan**
LOC Laye **EV.1SG-INAC** **être content**
'La nuit, au coucher, si je suis avec Laye, je suis **contente'**

(3d) **Omar** **ku** sedd **la** /* Omar **dafa-y** sedd
Omar **CL.REL** être froid **ECPT.3SG** / Omar **EV.3SG-INAC** être froid
'Omar est calme (tout le temps)', litt. 'Omar celui qui est froid'

[8] Mel'čuk rappelle une définition de l'opposition discret/continu: « Nous dirons que X est *linguistiquement continu* si et seulement si une partie quelconque de X s'appelle également X; autrement, X est *discret*. » (1994: 70).

L'exemple (3d) montre que la valeur de l'habitude de l'inaccompli *–y* n'est pas admissible. Le wolof utilisera de préférence le mode de l'emphatique du complément (ECPT) qui prend ici la valeur de l'existence et qui est nécessairement à l'accompli. La compatibilité de certains verbes d'émotions avec la forme d'inaccompli *–y* qui donne une valeur d'inchoatif, et avec le présentatif est remarquable (3e). Il faut noter que les verbes d'état de qualité n'acceptent pas le mode du présentatif (3f-g).

(3 e) **Dama-y** mer / **maa** **ngi** mer
EV.1SG-INAC être en colère / **1SG.MOD$_1$** **MOD$_2$** être en colère
'Je me fâche'

(3f) **Dafa** **tuuti** rekk de waaye **du** **ndaw**
EV.3SG **petit** seulement vraiment mais **EV.NEG.3SG** **jeune**
'Il est mince, mais il n'est pas jeune', litt. 'il est petit seulement vraiment mais il n'est pas jeune'

(3g) * **Mu** **ngi** tuuti rekk de waaye du ndaw
3SG.MOD1 **MOD2** être petit seulement vraiment mais EV.NEG.3SG être jeune

En wolof, les verbes d'émotion partagent avec les verbes d'état la possibilité d'être au mode de l'emphatique du verbe quand ils sont au présent (3b), alors que quelques uns d'entre eux partagent avec les verbes d'action la possibilité d'être au présentatif pour exprimer aussi le présent (3e).

Cette possibilité qu'ont des verbes d'émotion d'être tantôt employés comme état statif, tantôt comme état inchoatif a déjà été relevée par Apresjan (1992:70). Une forme inchoative dérivée de la base est d'ailleurs possible en wolof pour certains verbes d'état de qualité comme BAAX 'être bon' < BAAXSI 'devenir bon' alors que cette dérivation est impossible pour des verbes d'émotion comme MER 'être fâché' < *MERSI. Si cette dérivation a été relevée par S. Robert (1991 : 53) pour les verbes d'état de qualité, l'impossibilité pour les verbes d'état qui peuvent aussi avoir un sens de verbe d'action n'a pas été remarquée. Quelques verbes d'émotion comme MER 'être fâché, se fâcher', RAGAL 'avoir peur, prendre peur' possèdent la propriété combinatoire particulière d'être compatibles avec les aspects perfectifs et imperfectifs. Ils peuvent être utilisés comme des verbes d'état, ou comme des verbes d'action.

L'analogie de la classification aspectuelle des types de verbes et de la distinction noms comptables-noms massifs et a été reconnue en linguistique. J'y reviendrai dans ma dernière partie. On peut dès maintenant examiner cette distinction noms massifs-noms comptables en wolof et nous arrêter ensuite sur les noms d'émotion.

2 Une propriete combinatoire qui classifie des noms

2.1 Le comportement combinatoire des noms massifs et des noms comptables

Dans un article de 1998, Gillon a bien synthétisé les traits qui opposent les deux catégories des massifs et des comptables sur le plan morphosyntaxique. Les traits morphosyntaxiques mettent en avant le rôle des déterminants dans la distinction massif-comptable. Les langues cependant admettent des comportements différents vis à vis de ces critères: Kahane (2007) rappelle que les noms massifs abstraits en français se combinent avec le déterminant UN dès qu'ils sont modifiés mais qu'ils ne deviennent pas comptables pour autant. Il est donc nécessaire de savoir comment distinguer les noms massifs des noms comptables en wolof à l'aide de ces critères.

Il faut savoir qu'en wolof, tout lexème nominal dans un énoncé est associé à une marque de classe nominale. Il y a 8 classes au singulier et deux au pluriel : classes B, G, J, K, L, M, S, W au singulier et Y, Ñ au pluriel. Avant de pouvoir poser une hypothèse sur le statut massif ou comptable des noms d'émotion dans la langue wolof, il est nécessaire de savoir comment se comportent un nom de matière qui est un bon candidat pour être un nom massif, et un nom d'objet comptable. Considérons une matière liquide comme

l'eau, NDOX M- 'l'eau', et un objet comme un boubou (costume traditionnel en Afrique de l'ouest) : MBUBB M- 'le boubou'. Les exemples (4a-b) testent l'emploi des noms indéfinis :

(4a) Dama bëgge **ndox**
EV.1SG vouloir **eau**
'Je voudrais **de** l'eau', litt. 'je voudrais eau'

(4b) May na ko **(am / menn / ay)** mbubb
offrir PARF.3SG 3SG **INDEF.CL.SG / CL.SG.un / INDEF.CL.PL** boubou
'Il lui a offert (**un** indéfini**, un** numéral, **des**) boubou(s)'

Ces tests montrent que les noms massifs comme NDOX M- 'l'eau' ne prennent pas de marque de classe lorsqu'ils sont indéfinis, alors que les noms comptables nécessitent non seulement leur marque de classe mais également un déterminant indéfini. C'est une caractéristique intéressante des noms massifs en wolof. Pourtant, ces noms appartiennent bien à des classes nominales comme tous les autres noms. Dès qu'ils sont modifiés par un verbe de qualité comme en (4c) ou par un déterminant défini comme en (4d), le nom massif exige une marque de classe.

(4c) Dama bëgge **ndox mu** tàng
MOD.1SG vouloir **eau CL.REL** être chaud
'Je voudrais **de** l'eau chaude', litt. 'Je voudrais **eau** qui être chaude'

(4d) Dama bëgge **ndox mu** tàng **mi**
MOD.1SG vouloir **eau CL.REL** être chaud **CL.DEF**
'Je voudrais **l'**eau chaude', litt. 'Je voudrais **l'**eau **qui** être chaude'

Si l'on considère maintenant le cas d'un nom d'objet comptable, on remarque que la marque de classe est obligatoire dans tous les cas, que l'objet soit indéfini comme en (4b), défini comme en (5a), ou modifié par un verbe de qualité comme en (5b-c).

(5a) May na ko mbubb **mi / yi**
offrir PARF.3SG 3SG boubou **CL.DEF / CL.DEF**
'Il lui a offert **le/les** boubou(s)'

(5b) May na ko mbubb **mu** rafet **mi**
offrir PARF.3SG 3SG boubou **CL.REL** être beau **CL.DEF**
'Il lui a offert **le** beau boubou'

(5c) May na ko mbubb **yu** rafet **yi**
offrir PARF.3SG 3SG boubou **CL.REL** être beau **CL.DEF**
'Il lui a offert **les** beaux boubous'

L'autre observation que l'on peut tirer est que les noms massifs ne peuvent pas être utilisés au pluriel, que le nom massif soit indéfini comme en (6a) ou défini comme en (6b).

(6a) *Dama bëgge **ay** **ndox**
MOD.1SG vouloir **INDEF.CL.PL** **eau**

(6b) *Dama bëgge **ndox** (**yu** tàng) **yi**
MOD.1SG vouloir **eau** (**CL.REL** être chaud) **CL.PL.DEF**

J'ai relevé deux oppositions morphosyntaxiques entre les noms massifs et les noms comptables en wolof. Les noms massifs indéfinis ne prennent pas de marque de classe et ils ne se combinent pas avec la classe du pluriel *y-*. On peut se poser la question de la quantification pour les noms massifs. Nous y reviendrons lors de l'examen des noms d'émotion.

A présent que nous savons comment se comportent les noms massifs et les noms comptables en wolof relativement à la détermination, nous pouvons examiner le cas des noms d'émotion. Je me limite ici aux noms issus des bases verbo-nominales où le verbe est polysémique, c'est-à-dire des bases MER 'être en colère, se mettre en colère ; (M) colère' et RAGAL 'avoir peur, prendre peur ; (G) peur ; (B) peureux'.

2.2 Une propriete combinatoire des noms d'emotion

Les exemples (7a-b) montrent que ces noms d'émotion se comportent comme des noms massifs en wolof: le nom qui a le sens de 'émotion' en (7a) n'a pas de marque de classe à l'indéfini, alors que l'introduction d'un modifieur en (7b) nécessite une marque de classe.

(7a) Dama yég **ragal**
EV.1SG sentir **peur**
'Je ressens **de la peur'**, litt. 'Je ressens peur'

(7b) Dama yég **ragal** **gu** réy
EV.1SG sentir **peur** CL.REL être gros
'Je ressens **une grande peur'**, litt. 'Je ressens peur qui être gros'

Pourtant, l'emploi des numéraux avec les noms d'émotion est courant, comme en (8a-b), alors qu'il est impossible sans modification du sens avec des noms massifs de nourriture comme en (8c-d):

(8a) **Ñetti** **reccu** la ci am
Trois. PL **regret** ECPT.3SG LOC avoir
'Il a eu **trois** regrets (pour ça)', litt. 'Trois regrets il a dans ça'

(8b) Amuma ci **benn** jaaxle
Avoir.NEG.1SG LOC **un** inquiétude
'Je n'ai pas **une seule** inquiétude', litt. 'Je n'ai pas une inquiétude dans ça'

(8c) **Ceeb** laa togg / (8d) **Ñetti** **ceeb** laa togg
riz ECPT.1SG cuisiner / **trois** **riz** ECPT.1SG cuisiner
'J'ai préparé **du riz** au poisson' / 'J'ai préparé **trois portions** de riz au poisson'

Ces exemples illustrent la conversion de nom massif en nom comptable selon les procédés que rappellent Gillon (1998:) pour l'anglais et Kahane (2007: 7-9) pour le français.[9] Pour le wolof, la

[9] "Summarizing, then, we conclude that mass nouns, under conversion, give rise to count nouns with a limited variety of shifts in denotation. They include, but may not be confined to, the following: TO BE A KIND OF, TO BE AN INSTANCE OF, TO BE A UNIT OF, and TO BE A SOURCE OF": 57-58 ("Pour résumer donc, on conclut que les noms massifs donnent des noms comptables par conversion avec un nombre limité de glissements de sens. Sans se limiter à ce qui suit, cela inclut ETRE UNE SORTE DE, ETRE UN EXEMPLE DE, ETRE UNE PARTIE DE, et ETRE UNE SOURCE DE").

possibilité que certains noms d'émotion ont de se combiner avec des numéraux est corrélée avec un changement de sens. On peut noter cette différence de sens de la manière suivante (Kahane, 2007: 7): soit X un nom massif abstrait signifiant ‘émotion’, on pourra lui associer un sens signifiant ‘instance de l’émotion X’ sous la forme [$_{INST}$ X]. Cette compatibilité des noms d’émotion avec des déterminants des noms comptables fait partie de leurs propriétés combinatoires. Je peux définir cette propriété combinatoire: les noms d'émotion sont des noms massifs qui s'emploient au singulier sans marque de classe comme les autres noms massifs. Quand ils sont employés au pluriel, ils deviennent comptables et désignent non plus l’émotion comme telle, mais l’instance de l’émotion. Ils sont ainsi soumis à des conversions de massif en comptable. C'est cette disponibilité qui est une propriété combinatoire remarquable. Il n’est pas moins remarquable que le sens ‘instance de l'émotion X’ soit compatible avec des verbes de quantité (9a-b)

(9a) Seddam gi dafa **ëpp**
Calme.POSS.3SG CL.DEF EV.3SG **plus que**
‘Son calme est **exagere’**, litt. ‘Son calme est plus que’

(9b) Réccuwu ko lu **bare**
Regretter.NEG.3SG 3SG CL.REL **être nombreux**
‘Il ne le regrette pas **beaucoup',** litt. ‘Il ne regrette pas ce qui est nombreux’

Comme on peut le remarquer, les verbes de quantité eux aussi changent de sens. Quand ils sont combinés à des noms d'émotion, ils acquièrent le sens de l'intensité: le sens de BARE 'être nombreux' devient 'être beaucoup, être important, être d'une grande intensité'. Cette contrainte exercée sur le sens du quantifieur est tout à fait intéressante, mais elle mériterait une étude à part entière.

3. Vers une formalisation adaptee dans un modèle Sens-Texte

Les alternances de sens corrélées aux alternances combinatoires que j'ai relevées concernent aussi bien les verbes que les noms. L’analogie nom - verbe concerne l’opposition ‘discret / continu’ (structure interne de l’unité). La question qui se pose alors est: quelle représentation sémantique peut-on donner de ce phénomène dans un modèle Sens-Texte?

La sémantique conceptuelle de Jackendoff (1991) a proposé une formalisation particulièrement intéressante par l'ajout de traits (b pour ‘borné’ et i pour ‘structure interne’) aux structures lexicales conceptuelles (LCS) des lexèmes. Les verbes d’état et les noms massifs sont continus, donc dépourvus de structure interne. Ils reçoivent la valeur négative –i. De la même façon, ils ne sont pas bornés, donc ils reçoivent aussi la valeur négative –b. Les traits b / i sont ajoutés à la structure lexicale et conceptuelle du lexème qui se présente sous la forme d’une structure entre crochets. Des fonctions qui prennent les lexèmes munis de ces traits pour arguments permettent de changer les valeurs positives ou négatives initiales des traits. La fonction COMP (pour ‘composé de’) donne ainsi à une substance la valeur ‘quelque-chose composé de cette substance’. Jackendoff dit qu’elle représente la règle de l’ « universal packager » ou « emballage universel ». Bach et Jackendoff attribuent cette règle au philosophe Daniel Lewis dès les années 1960. Mais selon Jackendoff, l’intéressé lui-même nie en être l’auteur. Quoiqu’il en soit, cette règle représente celle que j’ai notée plus haut comme X → [$_{INST}$ X]. Elle change la valeur de la structure interne (le trait i) de l’unité linguistique. La fonction qui change la valeur du bornage (le trait b) est notée BD (pour ‘bounded’ en anglais). L'inchoatif peut ainsi être formalisé comme étant une fonction qui borne l'évènement par le commencement. Dans le cas qui nous concerne, l'inchoatif est une fonction qui prend pour argument un état et lui donne une valeur d'évènement culminant dans cet état.

On peut intégrer ce type de solutions dans un modèle Sens-Texte notamment dans la construction de patrons de polysémie (Barque, 2007: 29, pour l'utilisation de telles règles). Les patrons de polysémie sont des structures qui ont plusieurs champs d'informations. Ils se présentent sous forme de cadres imbriqués

les uns dans les autres (Barque, 2007: 93-101). Le premier cadre rassemble les informations liées à la sous-spécification sémantique. C'est ce champ qui est concerné par les propriétés combinatoires des bases verbo-nominales que nous avons dégagées. Nous proposons de rajouter ces informations dans ce cadre (figure 1 ci-dessous) en même temps que l'on procède à l'étiquetage sémantique des lexies. L'étiquetage sémantique est défini par Polguère (2003) comme une modélisation du sens de base d'une lexie dans un modèle Sens-Texte. Lorsque le sens d'une lexie est multiple, les étiquettes sémantiques sont combinées. On les note par « étiquette sémantique / étiquette sémantique ». J'illustre ci-dessous ma proposition avec la base verbo-nominale MER 'être en colère, se mettre en colère ; (M) colère' :
Etiquette sémantique de MER 'être en colère, se mettre en colère': **etat emotionnel /** action qui débute un **etat emotionnel**
Etiquette sémantique de MER M- 'colère' : **emotion** / manifestation d'une **emotion**
Deux règles permettent de décrire la polysémie des bases verbo-nominales. La règle de conversion '**etat emotionnel**' → **'emotion**' indique le lien entre les deux étiquettes sémantiques combinées (**etat emotionnel /** action qui débute un **etat emotionnel** : **emotion** / manifestation d'une **emotion**). Et la règle « d'emballage universel » décrit la combinaison « **emotion** / manifestation d'une **emotion** ». Dans le tableau, la règle de conversion se lit verticalement alors que la règle « d'emballage universel » se lit horizontalement.

Sous-specification semantique

Etiquette sémantique : état émotionnel	⟺	Etiquette sémantique : action qui débute un état émotionnel
⇕		
Etiquette sémantique : émotion	⟺	Etiquette sémantique : manifestation d'une émotion
Structure interne : continue	⟺	Structure interne : discrète

Règles lexicales semantiques

Conversion
« emballage universel»

Instances

MER 'être en colère, se mettre en colère'
MER (M) 'colère'

RAGAL 'avoir peur, prendre peur'
RAGAL (G) 'peur'

Figure 1. Cadre de sous-spécification dans un patron de polysémie

Conclusion

La langue wolof m'a permis de poursuivre une réflexion linguistique sur une analogie verbe d'état – nom massif déjà établie pour des langues mieux décrites comme l'anglais ou le français.

Robert (1991: 331-333) a établi une classification des types de procès pour le wolof et range dans une catégorie à part des verbes d'émotion comme RAGAL 'avoir peur, prendre peur', WAAR 'être étonné, s'étonner' ou encore des verbes de position du corps comme TOGG 'être assis, s'asseoir' en faisant remarquer que leur particularité est de pouvoir tantôt être utilisés comme des verbes d'état statifs, tantôt comme des verbes d'action dynamiques. S. Robert choisit pourtant de classer ce type de procès parmi les procès discrets, mais l'explication n'est pas clairement affirmée.

En m'inspirant de travaux de Jackendoff (1991), j'ai élargi cette réflexion pour le wolof en testant l'hypothèse que la compatibilité des verbes d'émotion avec l'inaccompli à valeur d'inchoatif est analogue à la compatibilité des noms d'émotion avec le pluriel. Il y a une polysémie de certaines bases verbo-nominales d'émotion due à leur particularité d'être soit continues et non bornées, soit discrètes et bornées.

Ce constat suscite une nouvelle question: quels moyens utiliser dans un modèle Sens-Texte pour formaliser ces règles ? Je propose d'intégrer les règles qui changent les valeurs des traits continu / discret dans les patrons de polysémie développés par Barque (2007).

Remerciements

Je tiens à remercier Felix K. Ameka, Jean-Léopold Diouf, et en particulier Sylvain Kahane pour leurs remarques et leurs corrections des versions antérieures.

References

Apresjan, Juri D. 1992. *Lexical semantics, user's guide to contemporary Russian vocabulary*. Ann Arbor, Karoma Publishers

Bach, Emmon. 1986. The Algebra of Events. *Linguistics and Philosophy,* n° 9, 5-16.

Barque, Lucie. 2007. *Description et formalisation de la polysémie régulière du français*. Université Paris-Diderot : s.n.

Comrie, Bernard. 1976. *Aspect: an introduction to the study of verbal aspect and related problems*. London, New-York, Melbourne: Cambridge University Press.

Diouf, Jean-Léopold. 2001. *Dictionnaire Wolof: wolof-français, français-wolof*. Tokyo: ILCAA.

Fal, Arame, Jean-L. Doneux, et Rosine Dos Santos. 1990. *Dictionnaire wolof-français*. Paris : Karthala.

Filip, Hana. Nominal and verbal semantic structure: analogies and interactions. *Language sciences*, Vol. 23, N° 4-5: 453-501, Elsevier.

Gillon, Brendan S. 1998. The lexical semantics of english count and mass nouns, in E. Viega (ed.) *Breadth and depth of semantic lexicon*.: 51-61, Kluwer Academic Publisher.

Jackendoff, Ray. 1991. Parts and boundaries. *Cognition*, Vol. 41: 9-45. Elsevier.

Kahane, Sylvain. 2007. La distribution des articles du français, in M. Charolles, N. Fournier, C. Fuchs & F. Lefeuvre (eds.), *Parcours de la phrase - Mélanges offerts à Pierre Le Goffic*: 159-174, Paris, Ophrys. www.kahane.fr/?u_act=download&dfile=Articles-2007.pdf&

Kleiber, Georges (ed.) 1989. *Termes massifs et termes comptables.* Actes du colloque des 26-27 novembre 1987 à l'université de Metz. Paris, Klincksieck.

Langacker, Ronald W. 1987. Nouns and verbs. *Language,* vol.63, n°1: 53-94. Linguistic Society of America.

Mel'čuk Igor. 1994-1996. *Cours de morphologie générale,* vol. 3 (1996) et vol.2 (1994). Montréal : PUM / CNRS

Mourelatos, A. P. D. 1978. Events, Processes, and States. *Linguistics and Philosophy,* n° 2: 415-434, Springer.

Nicolas, David. 2002. *La distinction entre noms massifs et noms comptables, aspects linguistiques et Conceptuels*. Louvain, Paris : Peeters.

Nouguier Voisin, Sylvie. 2002. *Relations entre fonctions syntaxiques et fonctions sémantiques en wolof.* Université de Lyon Louis Lumière: s.n.

Polguère, Alain. 2003. Étiquetage sémantique des lexies dans la base de données DiCo. *TAL, Traitement automatique des langues*, vol. 44, n° 2 : 39-68.

Robert, Stéphane. 2002. *Temps et verbe dans les langues africaines : l'exemple du wolof.* Notes du séminaire de DEA, 21 janvier.

Robert, Stéphane. 1998. Espace déictique, espace syntaxique et prédication: les indices spatiaux du wolof. In Bernard Caron (ed.) *Proceedings of the 16th International Congress of Linguists*, Paris: CNRS / Amsterdam: Elsevier.

Robert, Stéphane. 1991. *Approche énonciative du système verbal: le cas du Wolof.* Paris: CNRS.

Rothstein, Susan. 2008. Two puzzles for a theory of lexical aspect: semelfactives and degree achievements, in J. Dölling, T. Heyde-Zybatow, M. Schäfer (eds.) *Event Structures in Linguistic Form and Interpretation:* 175-198. Berlin, New-York: Walter de Gruyter.

Vendler, Zeno. 1967. *Linguistics in Philosophy*. New-York: Ithaca.

Korean Subject Attachment in Predicative Chains

Jihye Chun
MoDyCo, University of Paris Ouest Nanterre La Défense
chunjihye@gmail.com

Abstract

In this paper, we will examine the frequently observed chains of dependent Korean verbs sharing a single subject, and we will show that in some cases a low subject attachment seems to be the correct dependency analysis as it considerably simplifies subsequent linearization rules.

1 Linguistic Facts and Issues

Korean is known as a verb final language with relatively free word order for verbal arguments (Choi 1999, Chung 1995). Egedi *et al*.(1995) note that "a basic characteristic of Korean arguments is their ability to scramble, or move within the sentence". Furthermore, the subject placement is relatively free compared with other arguments.

First of all, let us consider the following examples in which two verbs share the same subject *Yeongi*:

(1) a. [sakwa-reul meok-kess-dako][1] **Yeongi-ka** eomma-hante[2] yaksokha-eoss-eo
apple-ACC eat-I-C[3] **Yeongi-SUBJ** mother-DAT promise-P-N[4]
'Yeongi promised her[5] mother to eat apples'

b. [sakwa-reul **Yeongi-ka** meok-kess-dako] eomma-hante yaksokha-eoss-eo
apple-ACC **Yeongi-SUBJ** eat-I-C mother-DAT promise-P-N
'Yeongi promised her mother to eat apples'

c. *[sakwa-reul Yeongi-ka **eomma-hante** meok-kess-dako] yaksokha-eoss-eo
apple-ACC Yeongi-SUBJ **mother-DAT** eat-I-C promise-P-N

d. *[Yeongi-ka meok-kess-dako] **sakwa-reul** eomma-hante yaksokha-eoss-eo
Yeongi-SUBJ eat-I-C **apple-ACC** mother-DAT promise-P-N

e. **eomma-hante** [sakwa-reul Yeongi-ka meok-kess-dako] yaksokha-eoss-eo
mother-DAT apple-ACC Yeongi-SUBJ eat-I-C promise-P-N
'Yeongi promised her mother to eat apples'

Note that in Korean grammar, the morphemes *-dako* and *-kess* are complementizers which introduce the embedded verb (Seo 1996). This is different from the infinitive in Indo-European Languages. Even

[1] The brackets are only intended to facilitate comprehension and the grammaticality judgments are independent of them.

[2] It exists also the dative marker *-eke*. However, in oral, we prefer to employ the marker *-hante*, while the marker *-eke* is more literal.

[3] I-C means respectively the intention and the complementizer. According to Seo (1996), this complementizer is used for expressing tense, that is, the present or the future. We will discuss this complementizer in the next section.

[4] P-N correspond respectively the past and the neutral form of verb.

[5] *eomma* 'mother' refers preferably to Yeongi's mother, but independently of the linear order between *Yeongi* and *eomma*, the word can also refer to the speaker's or somebody else's mother.

embedded, Korean verbs can have a tense and mood, and can realize all their dependents. Some complementizers, however, may restrict these possibilities (see example 3). There is no equivalent to an Indo-European infinitive.

These five sentences have the same semantic graph (predicative) in which the subject *Yeongi* is the first argument of *yaksokha-* 'promise' and that of *meok-* 'eat', and in which *eomma* 'mother' is an argument of *yaksokha-* 'promise' and *sakwa* 'apple' depends on *meok-* 'eat':

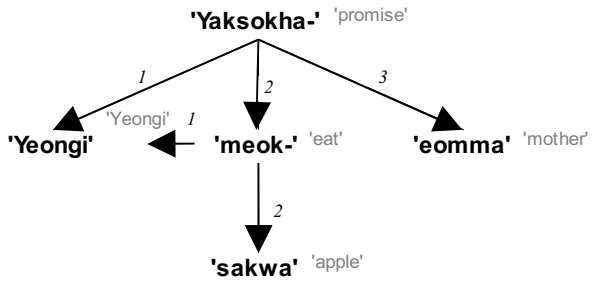


Figure I. Semantic graph of the sentences (1a-e)

First, let us compare the sentences (1a) and (1b): In the sentence (1a), the subject *Yeongi* is placed next to the embedded verb, that is to say, outside the bracket. However, in the sentence (1b), the subject occurs in front of the embedded verb, not the matrix verb. Furthermore, we believe that the communicative value of the sentence (1a) is different from that of the sentence (1b): In general, we produce the sentence (1a) in the context in which *Yeongi* hadn't wanted to eat apples, but her mother demanded that *Yeongi* eats apples perhaps for her health, and finally, *Yeongi* promised her mother to eat apples. On the other hand, the sentence (1b) emphasizes the fact that it's *Yeongi* who intends to eat apples. Note that we need to consider in this case prosodic factors such as intonation or pause.

The sentences (1c) and (1d) justify the argument dependence of each verb. Namely, the argument *sakwa* 'apple' of the verb *meok-* 'eat' has to be placed with its governor. This allows us to show that there is an embedded constituent into which no dependents of a higher verb can enter, for instance, the argument *eomma* 'mother' of the verb *yaksokha-* 'promise'. Note that in terms of the communicative structure, *eomma* 'mother' can be placed at the beginning of the sentence (see the sentence 1e). This means that as mentioned above, even though Korean is a free word order language, word order constraints do exist. Furthermore, we wonder why the overt subject *Yeongi* can be placed in front of the embedded verb *meok-* 'eat' in (1b) instead of being placed near the matrix verb. We suppose that free order of Korean subject creates ambiguity of subject attachment in the dependency tree. For this reason, we observe two theoretical possibilities: The subject could either attach to the embedded verb in the dependency tree, or depend on the matrix verb:

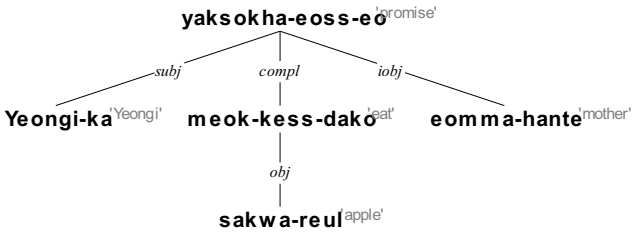


Figure II. Dependency tree I

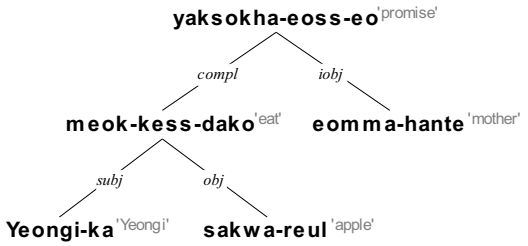


Figure III. Dependency tree II

This leads us to think about the general position of the subject in a dependency structure containing several verbs.

The central question we ask in this paper is how to find the correct dependency structure for the sentences (1a) and (1b). In European languages, the answer is clear in this case: If there is only one subject, it agrees with the only finite verb, and we suppose commonly that the subject depends on the higher verb. However, in Korean, we suppose that there are two possible analyses, in the first case it is a matter supposing that the two sentences have different dependency trees, (1a) high attachment (see Figure II) and (1b) low attachment (see Figure III); in the second case we suppose that this is only a topological problem and we will rely on the linearization rules to obtain the two possible word orders and we have only one syntactic structure in which the subject attaches to the embedded verb (see Figure III):

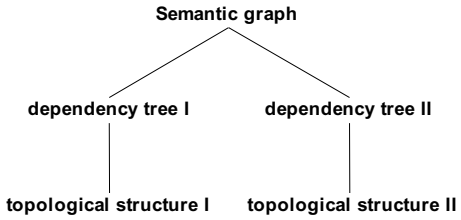


Figure IV. Analysis I

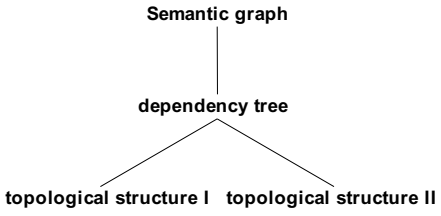


Figure V. Analysis II

We will show that two sentences have different syntactic structure and throughout this paper we will insist on the fact that with our examples, it is more convincing to adapt Analysis I. At first, we will argue that Analysis I is a convincing demonstration and then we will explain why we believe that Analysis I is more persuasive than Analysis II. We mainly compare our data with the Government and Binding Theory in terms of PRO, but we will remark that this framework cannot answer our questions in the Meaning-Text Theory framework and we suggest introducing further analysis, referring to Kong (1981) and to Roulet (2002). Finally, we will try to reformulate these observations within a dependency grammar.

With our main examples (1a) and (1b), it will become clear that the subject of the sentence could depend on the lowest verb under certain condition.

2 Behavior of Korean Subjects in Predicative Chains

In this section, we analyze subject attachment in predicative chains. In section 2.1., we will justify that, with the Government and Biding Theory, the subject depends on the matrix verb and the embedded verb has PRO. In section 2.2., following Kong (1981) and Roulet (2002), we will insist on the fact that it is the embedded verb which governs the subject. In each section, we try to reformulate in terms of dependency grammar: a description of the dependency tree and of the topological structure.

2.1 Subject Attachment to the Matrix Verb

Before observing examples in Korean, let us take an example in English:

(2) a. *Paul promised to eat apples*

b. *Paul promised* [PRO *to eat apples*]

Within the Government and Binding Theory, it is the matrix verb that has the overt subject, whereas the embedded verb has PRO, i.e. an empty pronoun in control cases. In the sentence (2a), the subject of the infinitive *eat* is controlled by the subject of the matrix verb *promise*. That is to say, the infinitive *eat* has its subject which is syntactically represented in the structure of this sentence, but not phonetically present, a so called PRO. Note that PRO is different from pro which is an empty subject for empty subject languages.

We suppose that the notion of PRO makes it possible to analyze the sentence (1a). However, we wonder if PRO has the same role in Korean, and although we will not discuss this issue in greater detail, but we believe that it is necessary to mention it briefly. According to Pak (2001) and Yang (1984), the idea of PRO should be modified and expanded so as to cover lexical or morphological causes of control such as the combinatory relationship between complementizer/mood marker and main verb in non-configurational languages like Korean (the sentence (la) is reproduced here as (3a) for convenience of the reader):

(3) a.[sakwa-reul meok-kess-dako] **Yeongi-ka** eomma-hante yaksokha-eoss-eo
apple-ACC eat-I-C **Yeongi-SUBJ** mother-DAT promise-P-N
'Yeongi promised her mother to eat apples'

b.[**(Cheolsu-ka)** sakwa-reul meok-eulkeo-lako] **Yeongi-ka** eomma-hante yaksokha-eoss-eo
(Cheolsu-SUBJ) apple-ACC eat-F-C **Yeongi-SUBJ** mother-DAT promise-P-N
'Yeongi promised her mother that Cheolsu would eat apple'
'Yeongi promised her mother that she would eat apples'

As shown by examples above, when the complementizer *-kess* is replaced by *-eulkeo*, there may be two subjects, namely, each verb has own subject. On the other hand, there must be only one subject in a sentence in which embedded verb combines with the complementizer *-kess*. According to Kim (2003), Korean linguists have different points of view[6] concerning the empty pronoun in embedded proposition. Following Moon (1989), Kim (2003) notes that the empty pronoun exists and is a case of pro, because the complementizer *-kess* is the morpheme which means its relation with the overt subject and it is this overt subject who has the intention or will to eat apples. Note that Yang (1984) emphasizes that the complementizer *-kess* can have two different meanings: intention or non-intention (in the latter case, it resembles future tense).

Now, let us come back to the following example, which is important for our analysis:

(4) [(EP)[7] sakwa-reul (EP) meok-kess-dako] **Yeongi-ka** eomma-hante yaksokha-eoss-eo
apple-ACC eat-I-C **Yeongi-SUBJ** mother-DAT promise-P-N
'Yeongi promised her mother to eat apples'

As we discussed above, like English example (2a), Korean subject *Yeongi* is a dependent of the matrix verb *yaksokha-* 'promise' and the embedded verb *meok-* 'eat' has PRO. It is justified by the matrix verb *yaksokha-* 'promise' and the presence of the complementizer *-kess*.

Now let us try to reformulate these observations within a dependency grammar and to describe the dependency tree of the sentence (1a) in which the subject *Yeongi* attaches to the matrix verb:

---

[6] According to Kim (2003), Moon (1989) emphasizes that this category is pro, while Lim (1987) signalizes that it is PRO. Furthermore, Yang (1984) suggests integrating PRO into pro for applying this terminology to Korean.

[7] Since subject can be placed freely in this completive, we mark EP 'empty pronoun' at a placement that the subject can occupy.

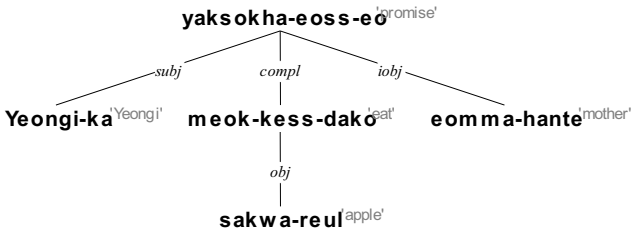


Figure VI. Dependency tree of the sentence (1a)

With this given dependency tree, we propose a Korean topological grammar following Gerdes & Kahane (2001). In our approach, placing an element in linear order means creating topological constituents (Gerdes & Kahane 2007): Each word can open a domain, containing a sequence of fields; these fields are possible positions for the (direct or indirect) dependents of the word. In the Korean topological grammar we propose, the main domain, opened by the highest verb, which is on top of the dependency tree, consists of the following sequence of two fields: main field and verbal field. The main field can accept any number of elements; the verbal field has to have exactly one occupant. Let us apply this grammar to our Korean example. The main verb *yaksokha-eoss-eo* 'promise' goes into the verbal field. As described by the dependency tree (Figure VI), the embedded verb *meok-* 'eat' cannot offer a placement for the subject inside completive box, while the subject *Yeongi* is placed in the main field opened by the matrix verb *yaksokha-eoss-eo* 'promise':

Main field			Verbal field
Completive domain		**Yeongi-ka eomma-hante**	*Verb cluster*
sakwa-reul	**meok-kess-dako**		**yaksokha-eoss-eo**

Figure VII. Topological structure of the sentence (1a)

Gerdes & Kahane (2007) note that the dependents of a verb do not have to be in their governor's domain: They can be "emancipated" and end up in a superior domain. However, as shown by the sentence (1d), the dependent *sakwa* 'apple' cannot emancipate its governor *meok-* 'eat's domain. This indicates that if we make the assumption that the subject *Yeongi* depends on the embedded verb (see Figure III), *Yeongi* cannot emancipate its governor, either. That is why we think that Analysis II doesn't cover the sentence (1a). In this case, we tend to produce the sentence (1b) in which the subject occurs in front of the embedded verb. Therefore, we consider that in the dependency tree of the sentence (1a), the subject attaches to the matrix verb and the emancipation of the subject *Yeongi* doesn't exist.

In the following section, we analyze our second example (1b) and we propose another syntactic structure and topological structure in comparison to those of the sentence (1a).

2.2 Subject Attachment to the Embedded Verb

First of all, let us observe the following example (the sentence (lb) is reproduced here as (5) for convenience of the reader):

(5) [sakwa-reul **Yeongi-ka** meok-kess-dako] eomma-hante yaksokha-eoss-eo
apple-ACC **Yeongi-ka** eat-I-C mother-DAT promise-P-N
'Yeongi promised her mother to eat apples'

We see that this sentence doesn't verify the Binding Principal which we discussed in the previous section because the subject in this sentence behaves like that of the embedded verb *eat* in the completive. That's why we have difficulties to explain this sentence with the Government and Binding Theory. However, there are authors who remark that concerning this type of sentences there are limits from a syntactical point of view, and they suggest analyzing this type of sentences in a cognitive or discursive approach. We

attempt to analyze this sentence, following Kong (1981) and Roulet (2002).

Kong (1981) suggests a perceptual analysis to solve this problem that we were not able to explain in a syntactical approach (a "static model", according to him), by insisting that perceptual or cognitive unit has to be used as a co-referential domain. This means that when we interpret sentences and analyze pronoun rules, it is more convincing that a perceptual unit behaves as the base unit. His analysis starts from the following example which violates the Binding Principal:

(6) *In the bed which* ***Zelda*** (i) *stole from the Salvation Army,* ***she*** (i) *spent her sweetest hours*
(Kong 1981:101)

He believes that, after a perceptual unit finishes treating information, this information is transferred to next level, so that we can minimize loss of the memory capacity. He develops his argumentation by saying that if this sentence is in a governed domain, we cannot have co-referential relation in terms of the Binding Principal. However, if we consider that a detached prepositional phrase is a perceptually independent unit, the analysis becomes more convincing: If we suppose that the sentence (6) is only one perceptual unit, we cannot constitute co-referential relation between *Zelda* and *she*. On the other hand, if we follow the idea of Kong (1981), this sentence consists of two perceptual units, therefore *Zelda* and *she* belong to another unit and we don't have any problems to constitute co-referential relation between *Zelda* and *she*. That is to say, perceptual segmentation allows us to capture what the Binding Principal cannot explain as phenomena.

It seems to us that the idea of Kong (1981) corresponds to that of Roulet (2002). At first, let us observe the following examples:

(7) a. *J'ai téléphoné à* ***la voisine*** (i) *pour que* ***la brave femme*** (i)[8] *m'achète du thé*
'I called the neighbor (i) so that the friendly lady (i) would buy me some tea'

b. ***Mon voisin*** (i) *m'a dit qu'****il*** (i) *(*****le pauvre homme****) était malade*
'My neighbor (i) told me that he (i) (*the poor man) was ill'
(Roulet 2002:168)

Roulet (2002) explains the sentence (7a) in terms of discourse memory and that only one passage by discourse memory allows establishing a co-referential link between *la voisine* and *la brave femme*. Therefore, the sentence (7a) must be analyzed as containing two discourse acts. However, in the sentence (7b), a co-referential link cannot be established between *mon voisin* and *le pauvre homme*. Because transfer to discourse memory has not yet taken place for this segment, thus, the sentence (7b) is analyzed as only one act.

Following Kong (1981) and Roulet (2002), we analyze the sentence (5) as two independent discourse segmentations: The first segmentation is 'Yeongi has the intention to eat apples' and the second segmentation is 'she promised that'; a co-referential link can be established between *Yeongi* and *she* in terms of discourse memory or perceptual units. Moreover, prosodically, this sentence is very interesting because in general, Korean speakers need to pause after producing a completive phrase, showing that this completive is detached from the principal proposition:

(8) [sakwa-reul **Yeongi-ka** meok-kess-dako] / (EP) eomma-hante (EP) yaksokha-eoss-eo
apple-ACC **Yeongi-SUBJ** eat-I-C mother-DAT promise-P-N
'Yeongi promised her mother to eat apples'

Moreover, Korean is a pro-drop language and subjects can be omitted, so it is not surprising that the subject is not expressed, sometimes, it is more natural to omit the subject.[9]

In general, a completive is governed by the principal proposition. However, we remark that Korean

[8] *la brave femme* obviously can be marked by the pronoun *elle*.

[9] According Huang (1984), languages such as Chinese and Korean so called discourse-oriented languages that entirely lack agreements, allow null arguments both in the subject and object position; whenever they pick up antecedents from discourse.

completive behaves differently compared to English or French. For example, it is not possible that a French completive is detached from principal proposition:

(9) a. ?? *Qu'elle mangerait une pomme, Marie l' a promis*
that she would_eat an apple Marie it has promised
'that she would eat an apple, Marie promised it'

b. ?* *Que Marie mangerait une pomme, elle l' a promis*
that Marie would_eat an apple she it has promised
'that Marie would eat an apple, she promised it'

Grobet (1997) notes that the prosodic criterion must be taken into account only when other criteria fail to explain the phenomena.[10] A prosodic approach strengthens the thesis of Kong (1981) and that of Roulet (2002) and the fact that the sentence (1b) consists of two independent segmentations. We believe that the dependency tree of the sentence (1b) is different from the dependency tree of the sentence (1a) and that the subject of the sentence (1b) depends on the embedded verb:

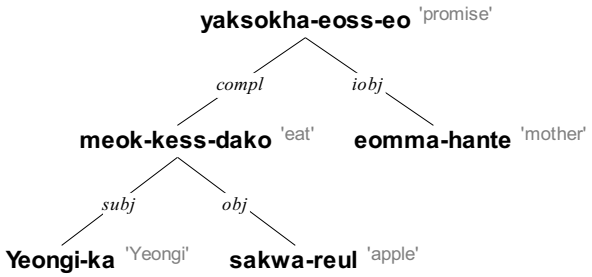


Figure VIII. Dependency tree of the sentence (1b)

Contrary to the topological structure of the sentence (1a), the subject *Yeongi-ka* is placed in the completive box opened by its governor *meok-* 'eat', and as shown by (1d), the dependents of the embedded verb *meok-* 'eat' cannot emancipate itself from its governor's domain:

Main field		Verbal field
Completive domain	**eomma-hante**	*Verb cluster*
sakwa-reul Yeongi-ka | **meok-kess-dako**		**yaksokha-eoss-eo**

Figure IX. Topological structure of the sentence (1b)

In a syntax-topology interface, we consider that it is impossible that the highest word goes into the embedded box; moreover, in the sentence (1b), the subject is placed in front of the embedded verb instead of being placed near the matrix verb. Therefore, we might suppose that in principle the subject *Yeongi* attaches to the embedded verb as described by the dependency tree above (see Figure VIII) and as Analysis II (see Figure V). However, if we suppose that this is only a topological problem and the sentences (1a) and (1b) have same syntactic structure in which the subject depends on the embedded verb, we don't have any argument to explain the sentence (1a) because as we have discussed, we believe that the subject *Yeongi* cannot emancipate its governor, for example *sakwa* 'apple' (see sentence 1d), which is to say, *Yeongi* is always placed in the completive domain. Therefore, we think that the sentences (1a) and (1b) have different syntactic structure, and this observation permits to think that we are able to constitute a dependency tree and a topological structure at the same time for these two sentences.

[10] « Il me semble que le critère prosodique doit être pris en compte uniquement quand les autres critères font défaut. »

3 Conclusion

We have discussed the position of Korean subject, especially comparing the sentences (1a) and (1b). Throughout this paper, we defend the idea that, if we have several verbs in the sentence that share the same agent, the subject in Korean doesn't always attach to the highest verb in a dependency tree, because in some cases, we have a detached completive phrase that has its own subject.

Acknowledgements

I would like to thank my dear professors Sylvain Kahane and Kim Gerdes who gave me important criticisms, remarks and suggestions. This paper would not have been possible without them.

References

Choi, Hye-Won. 1999. *Optimizing Structure in Context: Scrambling and Information Structure*, Stanford: CSLI Publications.

Chung, Chan. 1995. A *Lexical Approach to Word Order Variation in Korean*, Ph.D. Dissertation, Ohio State University.

Gerdes, Kim., & Kahane, Sylvain. 2001. Word order in German: a formal dependency grammar using a topological hierarchy, *Proceedings Association for Computational Linguistics 2001*, Toulouse.

Gerdes, Kim., & Kahane, Sylvain. 2007. Phrasing it differently, in L. Wanner(ed.), *Selected lexical and grammatical issues in the Meaning-Text Theory*, pp.297-335.

Grobet, Anne. 1997. La punctuation prosodique dans les dimensions périodiques et informationnelles du discours, *Cahiers de linguistique française*,19, pp.83-123.

Heycock, Caroline., & Lee, Young-Suk. 1989. Subjects and predication in Korean and Japanese, *Japanese/Korean Linguistics*, Vol.2, pp.239-254.

Huang, C.-T. James. 1984. On the distribution and Reference of Empty Pronouns, *Linguistic Inquiry*,15, pp.531-574.

Jang, Youngjun. 2002. Small Clauses and Default Case, *Proceedings of the 16th Pacific Asia Conference*, Jeju, Korea.

Kim, Mi-Young. 2003. *An Optimality Approach to the Referential Interpretation of Zero anaphora in Korean*, Ph.D. Dissertation, Seoul National University.

Kim, So-Young. 2007. Topics and Null arguments in Korean: the syntax and discourse, *Proceedings of Workshop in General Linguistics 2006*, pp.63-76.

Kong, Young-Il. 1981. Daemyeongsawa Munjangeui Inji (Pronoun and perceptual sentence), *Eoneowa eoneohak (Language and Linguistics)*, 7, pp.91-109.

Mel'cŭk, Igor A. 1988. *Dependency Syntax*, New York: State University of New York Press.

Pak, Duk-Soo. 2001. Lexical local control in Korean, In Y.-A Cho (ed.), *Korean studies at the dawn of the millennium: Proceedings of the second biennial conference, Korean studies Association of Australasia*, pp.276-286.

Egedi, Dania., Palmer, Martha., & Park, Hyun. S. 1995. Recovering Empty Arguments in Korean, *Proceeding of the 1994 Kyoto Conference*, pp.73-82.

Park, You-Jeong. 1990. *L'ellipse du sujet grammatical en coréen contemporain: étude comprative avec le français*, Thèse de doctorat, Université Paris Sorbonne.

Roulet, Eddy. 2002. Le problem de la définition des unités à la frontière entre le syntaxique et le textuel, *Verbum* 24, pp.161-178.

Seo, Jeong-Su. 1996. Kukeo Munbeop (*Korean grammar*), Seoul: Hanyang University Press.

Yang, Dong-Whee. 1984. Hwakdae Tongje Iron (Extended Control Theory), *Language Research*, Vol. 20, pp.19-30.

Yang, Dong-Whee. 1985. Kyeolsok Jibae Ironeui Ironjeok Baekyeong (Theoretical background of the government and binding theory), *Eoneowa eoneohak (Language and Linguistics)*, 11, pp.7-44.

Zribi-Hertz, Anne. 1998. *L'anaphore et les pronoms: une introduction à la syntaxe générative*, Paris: Septentrion.

Sharing the Knowledge of Lexicographers: Methodology for the Extraction of Lexicographic Abilities

Sophie Comeau
OLST—Département de linguistique et de traduction
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal (QC) H3C 3J7
tel. (1-514) 343-6111 Ext. 32585
fax (1-514) 343-2284
Sophie.comeau@umontreal.ca

Abstract

In our paper, we will present the Lexitation project, which is, to our knowledge, the first attempt at extracting lexicographic abilities using experimental techniques. We will describe the methods and results of our initial set of experiments, that are based on the use of so-called Think Aloud Protocol (Ericsson & Simon, 1993). We will explain how experiments have been set up and how we are currently proceeding with the extraction and modeling of various types of knowledge and strategies used by lexicographers while performing lexicographic tasks. Finally, we will present possible applications of our work in the field of language teaching.

1 Introduction

Lexicography is a discipline that has been extensively documented (Zgusta, 1971; Béjoint, 2000; Landau, 2001, Mel'čuk 2006). Lately, a few authors have been discussing the relation between some of these writings, seen as "theoretical", and the more "practical" field of dictionary-making (Fontenelle, 2008; Atkins & Rundell, 2008). However, there exists no systematic study of lexicographic abilities—i.e. declarative and procedural knowledge mastered by trained lexicographers—and no structured model of descriptive and methodological concepts involved in lexicography is available. Hence, the work we are presenting here targets the building of an ontological model of lexicographic abilities. It is part of a four-year project called Ontologization of lexicographic abilities for use in the fields of applied linguistics, nicknamed Lexitation, from which pedagogical applications will be derived. The project relies on the assumption that lexicographic abilities play a role in teaching and acquisition of lexical knowledge, and not only in lexicography per se. The originality of this research is that we extract lexicographic abilities using experimental techniques. Basically, we invite professional lexicographers to come and work before our eyes, and we record the work sessions. The recordings are later scrutinized to extract lexicographic abilities that we model into an ontology designed with the Protégé ontology editor.[1] Although our background is that of Explanatory Combinatorial Lexicology (Mel'čuk et al., 1995; Mel'čuk, 2006), we target the modeling of concepts and strategies used by lexicographers regardless of their specific theoretical background. We are still at the beginning of the project : five experiments took place so far. In the next sections, we will present results of the first stages of knowledge extraction from Experiment 3.

[1]Protégé is a free ontology editor and knowledge-base framework, developed at Stanford University in collaboration with the University of Manchester.

2 Description of experimental approach

2.1 Experimental settings

The goal of our experiments is to extract the knowledge and know-how involved in lexicography. To do so, we need to observe some lexicographers at work. Hence, for each experiment, we invite a lexicographer to come to a laboratory where he or she performs a lexicographic activity that we record. Five experiments took place so far. The first two were used as tests; participants were students trained for Explanatory Combinatorial Lexicology. The other three participants are professional lexicographers, two for academic dictionaries and one for a private company. Two of them are from Québec and one is from France. We used a verbal protocol to gain information about their cognitive thoughts while performing a lexicographic task. Ericsson and Simon (1984) present concurrent (also referred to as Think Alouds) and retrospective (also referred to as Think Afters) verbal protocols as ways to generate data about cognitive processes. In concurrent verbal reports, participants are asked to report their thoughts during the performance of a task, whereas retrospective verbal reports rely on gathering information after a task is carried out. Although both techniques can be influenced by a motivational shift, that can occur whenever participants are informed they are being observed (Russo, Johnson & Stephens, 1989), they provide a great amount of data about cognitive and behavioral processes. Though the Think Aloud protocol can be a little invasive, it was chosen over the Think After protocol, because the latter is more likely to be influenced by forgetting and fabrication, as shown by Branch (2000) and Ericsson and Simon (1984). Moreover, we found out through our first experiments that the more the participants are experienced, the less they have difficulty in doing Think Alouds. This is consistent with the studies of Ericsson and Simon (1984) and of Branch (2000), who indicated that Think Alouds are less useful when participants are carrying out a new task or a task that involves a high cognitive load. Our participants all being experienced professional lexicographers (except for the test phase), they seemed not too disturbed by the Think Aloud protocol. Hence, we prefer to keep using that protocol, thus avoiding the problem of incomplete memories that can occur with the Think Afters. In the experiments, to monitor that procedure, we had a research assistant we called the “pilot” to sit next to the lexicographers and to encourage them to think out loud, by asking them questions such as “Why did you do that ?” or “Can you tell me what you are thinking at the very moment ?” etc. The pilot would also help them out whenever they had a technical problem. While performing the task, participants were audio and video taped by 3 cameras, and we had the computer screen recorded too. The lexicographer and the pilot would sit in a room, the workstation, while the “controllers” – i.e. other experimenters – would watch the recordings on TV screens from another room. They could see and hear what was going on in the workstation in real time, so they could address comments or instructions to the pilot through a microphone connected to the pilot’s headphone.

2.2 Lexicographic activities

The same lexicographic activity (see definition in 3.2) was used for the first three experiments. Participants were given the assignment to outline the polysemic structure of French vocable COMPTER, ‘to count’, that is, finding its different senses. The computer used was provided with French corpus Le Migou, build from French newspaper *Le Monde*, and participants could access the Web via a browser. They were told they could use any of the two corpora, or both, but that there was no obligation to. They were asked to work accordingly to their usual methods as far as possible. They were given up to an hour to fulfill the task, but we insisted it could be less if they didn't need as much time. They had to store each lexical unit they found in a separate file of a document, which model we had created previously with Filemaker software. They could create as many files as they needed. Each file form contained five fields for them to fill in:

1. LEXICAL UNIT, where they would write down the lexical unit’s name and a number if they wanted to (though they were not instructed to).

2. GRAMMATICAL CHARACTERISTICS, where they could write the lexical unit's part of speech, or any characteristic they found worthy of noticing.

3. CLUE, where they could write something that would allow them to distinguish the lexical unit from others. It could be a quasi-synonym, for example.

4. EXAMPLE, where they would write as an example a sentence containing the lexical unit. It could be their own example or it could be taken from corpus.

5. COMMENTS, where they could write any comment they found relevant.

That was the activity for Experiments 1 to 3. For Experiments 4 and 5, though, we asked the lexicographers to do, for an hour or so, the typical work they are trained to do, and that they actually do in their everyday job, i.e. writing or editing dictionary articles. They would therefore work on their own documents, using their own tools. Actually, we favor activities that simulate the usual lexicographic work of our participants, the objective being to benefit as much as possible from their participation and to extract as much lexicographic abilities as possible.

3 Methodology for data analysis

The extraction of lexicographic abilities that will be described here was made from data of Experiment 3 only. Data of other experiments are being currently analyzed.

3.1 Inductive techniques

Our approach to data analysis fits the category Ericsson & Simon (1984) call "meaningful analysis of verbalizations", where there is no agreement, between the experimenter and the subject, upon specific signals. In this type of analysis, there are at least two ways to proceed. One way is to analyze the data in terms of their meanings, with a theory guiding the analysis, which limits the encoding to selected aspects and features and, when encoding the data, to map the verbalizations onto these categories of concepts and features. Our approach to extraction does not fit the above scheme. Our method was rather inductive, that is, we did not try to analyze the data along with a set of predetermined classes of lexicographic abilities. Indeed, we would rather scrutinize the "raw data" and try to identify what knowledge and strategies were used by the lexicographer along the experiment, and only then we would create corresponding abstract classes of lexicographic abilities. This procedure is perfectly legitimate, according to Ericsson & Simon (1989:6): "In less formal kinds of analysis, the encoding scheme is not defined formally and a priori, but the search for interpretations proceeds in parallel with the search for an appropriate model or theory. We recognize clearly the need for and value of such interactive processes in the search for theories in new domains". Lexicography is not what we can call a "new domain" ; Dr. Johnson addressed his "Plan of a Dictionary" to Lord Chesterfield in 1747 and, since then, many authors have set theoretical foundations for the discipline (Zgusta, 1971; Béjoint, 2000; Landau, 2001; Mel'čuk et al., 1995; Mel'čuk, 2006, to name a few). However, there exist no structured model of descriptive and methodological concepts and abilities involved in lexicography. Therefore, when analyzing the data, we could not categorize each verbalization into one existing category. Although we plan to enrich our ontology with concepts taken from the literature once the analysis is over, we decided to start from the observation of the experiments to create categories of lexicographic abilities as they appear. As a result of this choice, we had to define our own methodology for data analysis and our own encoding scheme. We'll explain how we proceeded in the next subsections.

3.2 Some definitions

Data analysis relies on three basic concepts. It is necessary to define them before to go any further with the description of the methodology:

LEXICOGRAPHIC ACTIVITY: The assignment given to the lexicographer for a given experiment. The activity is usually chosen and designed by the team of experimenters, but it can also be, like it was the case for Experiments 4 and 5, an activity taken from the lexicographer's daily duty. There is only one lexicographic activity per experiment.

LEXICOGRAPHIC TASK: Segment of the activity associated with a given time period in which the lexicographer accomplishes a specific task. What is important here is that the segmentation of the activity into tasks is a linear one; each task is given a number and its starting and ending time is identified. For Experiment 3, that lasted 55 minutes, we identified a total of 40 tasks, that means that tasks have a mean duration of nearly a minute and a half.

LEXICOGRAPHIC OPERATION: "minimal" action taken by the lexicographer inside a given task. Operations are not associated with given time sequences. We have to explain here that operations correspond to very short sequences that can overlap. They can also be inferred, that is, we can assume that the lexicographer does one particular operation even though he doesn't say so – If there are some evidences that allow us to believe so –. For it was difficult, if not impossible, at times, to locate operations in time, we decided to order them logically rather than chronologically.

The three concepts introduced here will be illustrated in next sections. Let us not forget that the analysis presented in this text was made with data from only one experiment; we expect various different types of tasks and operations to add up as other experiments are being analyzed.

3.3 Division of activity into tasks

After seeing recordings of Experiment 3, we first divided the footage in 40 great "scenes", corresponding to the tasks accomplished by the lexicographer. These 40 tasks were mainly of three different types, according to the goal the lexicographer pursues in doing them. In our ontology, we created three abstract classes of tasks, corresponding to the three types of tasks we sorted out, and each of the 40 tasks is considered an instance of one of the three classes, which will be described here:

• SEARCH FOR SENSE: The lexicographer is searching for senses he haven't thought of yet, using the corpus or by introspection.

• CONCEPTUALIZATION OF SENSE: The lexicographer conceptualizes a lexical unit and writes a short lexicographic description in a file.

• MODIFICATION OF DESCRIPTION: The lexicographer operates a minor modification in one of his files.

We decided to include in the class MODIFICATION OF DESCRIPTION only the instances of tasks in which the modification does not imply a reorganization of the structure of the vocable as it is outlined by the lexicographer at the time. In general, a modification aims at improving a lexicographic description, by making it more complete or by conforming to dictionary-writing rules. For example, in task 18 of Experiment 3, the lexicographer adds an example taken from a corpus in his file COMPTER-'dénombrer', that already had an example made up by him. He adds it because an example taken from "real speech" would confer more legitimacy to his description. In this case, there is no conceptualization of a lexical unit, but a "superficial" modification in the lexicographic description.

3.4 Division of tasks into operations

At this stage of the analysis, we had in the ontology classes of lexicographic tasks and instances of these classes. Next step was to divide instances of tasks into simpler actions, called "operations", and create abstract classes of operations, just like we did for the tasks. These two groups of classes are independents.

Classes of operations are not subclasses of tasks; the same type of operations can be performed in different types of tasks. Here are the classes of operations identified to date, and their hierarchy:

Note the following abbreviations:

LU means 'lexical unit'; QSYN means 'quasi-synonym'; ASSESS means 'assessment'; EX means 'example'; POS means 'part of speech'; SPECIF means 'specification'.

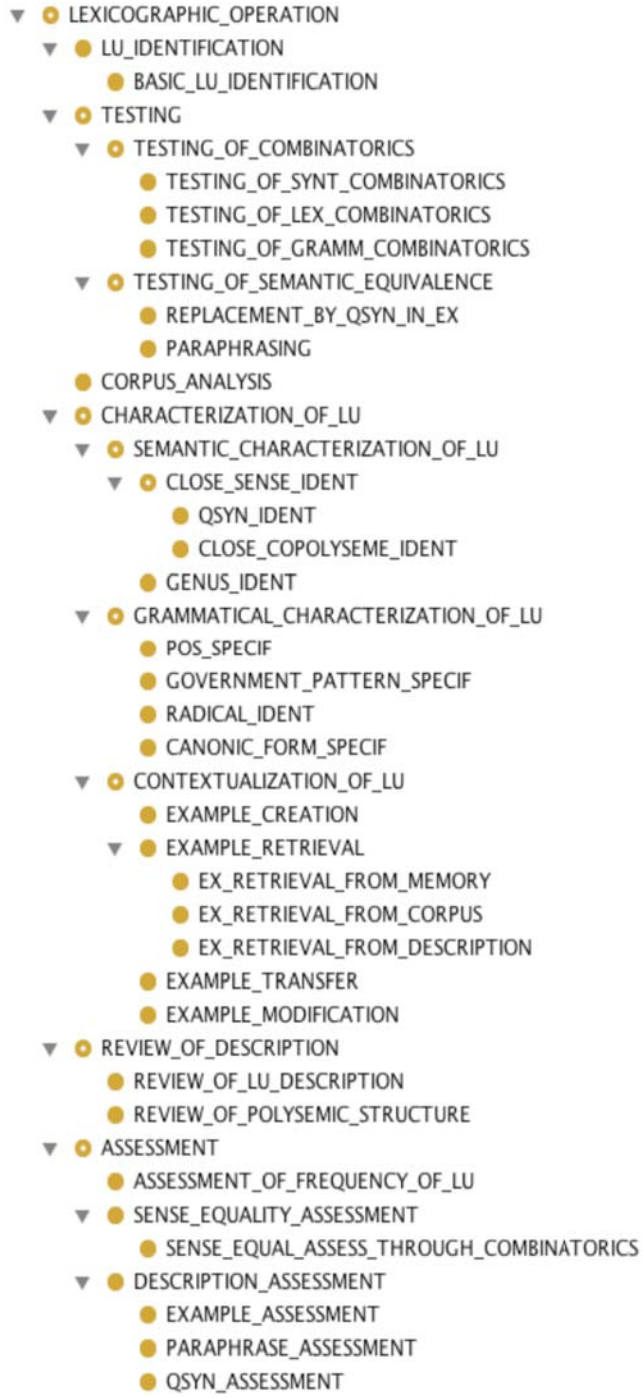


Figure 1. Hierarchy of lexicographic operations

Figure 1 displays all the classes and subclasses of operations that we observed in one or more tasks of the Experiment 3. In each specific task, the participant carries out some of these operations, but not necessarily in the order they are presented in the hierarchy. Classes of operations themselves are independent from one another. For example, the classes TESTING and LU_IDENTIFICATION are in a sister-sister relation rather than in a mother-daughter one, but BASIC_LU_IDENTIFICATION is a subclass (daughter) of the class LU_IDENTIFICATION. Other classes of operations will be added to these as data from other experiments are analyzed. As we mentioned before, operations are sometimes inferred, that is, we assume that the lexicographer carries them out even though he doesn't mention them. Of course, we try to infer as less as possible, but enough to obtain a complete and logic chain of operations. Most times, inferences rely on tangible evidences. For example, in task 4 of Experiment 3, lexicographer writes down the part of speech of the lexical unit he is thinking of, a verb, as he is talking about something else. Though he does not say "I am now identifying the part of speech of the lexical unit", we can infer he did and add POS_SPECIF to the list of operations of task 4. But since operations themselves are sometimes not tangibles, it is often impossible to identify exactly at what time they occur. Moreover, they are of very short duration and can often overlap. In this same task 4, for example, when the lexicographer is writing down the part of speech of the lexical unit, he is also uttering the hypothesis of this sense being distinct from another. The two operations take place simultaneously, and this is a pretty common case. Therefore, we decided not to identify time sequence of each operation, like we did for tasks. But what we can do is to order the operations according to an ordinal and logical order. Indeed, in the ontology, in each instance of operation, we identify the operations that led to that one. To make the methodology of the extraction more clear, let's see the transcript of a task taken from Experiment 3 and what classes of abilities we extracted from it.

4 Transcript of a task and its analysis

Here is the transcript of task 8 of Experiment 3. Italic is used to cite the lexicographer; translation follows, in parenthesis. Words in small capital letters are the names of the classes of operations we created.

1. « *Ça me fait penser, hier, je jouais avec ma nièce...elle a dit : " Ça, ça compte pas."* » (This reminds me, yesterday, I was playing with my niece...She said "that doesn't count").
= EX_RETRIEVAL_FROM_MEMORY

2. « *Donc... "C'est pas conforme au règlement."* » (So... "It doesn't go by the rule").
= QSYN_IDENTIFICATION

3. « *Ou... Il faut pas en tenir compte … Il faut pas prendre en consideration… Ça ressemble un petit peu mais c'est pas "compter avec", là…*» (Or... "We shouldn't take it into account... We shouldn't take it into consideration… It's a bit similar, but it's not "compter avec").
= LU_IDENTIFICATION

First, in this operation, the lexicographer compares the sense of his retrieved example, *«Ça compte pas»*, with the sense of another lexical unit of the same vocable, semantically close, that he tagged *COMPTER AVEC*- 'prendre en considération, tenir compte de'. (TO COUNT 'take into consideration'). Then, he states that the meaning of the lexical unit he retrieved is different from that of *COMPTER AVEC*, and he decides that this particular sense is worthy of the lexical unit status, so he creates a new file.

4. He identifies and writes down the part of speech of the lexical unit.
= POS_SPECIF

5. He writes down the quasi-synonym "être officiel".
= QSYN_IDENTIFICATION

6. He writes down another quasi-synonym, “réglementaire”.
= QSYN_IDENTIFICATION

7. He identifies another quasi-synonym, “homologué”.
= QSYN_IDENTIFICATION

8. « *Cette ronde de pratique ne compte pas.* » (This turn does not count).
= EXAMPLE_CREATION.

Here, he invents a sentence and writes it as an example.

9.« *Ça, ça compte pas.* » (That does not count).
= EX_RETRIEVAL_FROM_MEMORY

Here he recalls the same example he heard the day before, and he writes it down.

10. « *C’est peut-être un peu trop relâché.* » (This is maybe too informal).
= EXAMPLE_ ASSESSMENT.

11. « *Je suis pas entièrement satisfait de ma description...Il y aurait peut-être quelque chose de plus juste.*» (I’m not entirely satisfied with my description...There could be something more accurate).
= QSYN_ ASSESSMENT.

Here, he stares at his file, and when asked what he is thinking about, he says he is not entirely satisfied with his description, pointing at the quasi-synonym. Finally, he passes on to something else, saying that the quasi-synonyms are mainly there to help distinguish one lexical unit from the others.

We’ll see now how classes of abilities are organized and encoded in the ontology.

5 Encoding of data

The encoding of data was done in an ontology designed with Protégé ontology editor. An ontology is a formal explicit description of concepts (classes) in a domain of discourse – in this case, lexicography –, properties of each concept describing various features and attributes of the concept (slots), and a set of individual instances of classes. Classes describe concepts in the domain, and slots describe properties of classes and instances (Noy & McGuinness, 2001). In our ontology, concepts correspond to the classes of lexicographic tasks and operations (as shown in Figure 1). Tasks and operations performed during the experiments are instances of these classes. In the ontology, there is a window for each instance of task or operation, in which various slots display information about the instance itself. For example, in each instance of task, we display its starting and ending time, the video of the task, and all the lexicological concepts used by the lexicographer during the task.

6 Pedagogical applications

As we said before, Lexitation project is in its beginnings. We are planning other experiments, some of which with English lexicographers. Experiments 4 and 5 will soon be analyzed, according to the method we exposed in this text. That being said, there is one question left to answer: «What purposes will the ontology serve?» We think it could be very useful, among other things, in language teaching, especially for teacher training. In the literature about language teaching, it is a pretty common place to say that lexicon is not given enough attention and that more work should be done in this field. In fact, authors from Québec (Simard, 1994), Switzerland (De Pietro, 2003) and France (Grossmann et al., 2005) deplore the lack of a systematic teaching of lexicon that would accompany the natural and spontaneous interventions about lexical features that teachers make on a daily basis as reading or other types of

activities are done in class. However, we think that the potential of our project lies right there, in the interventions about lexicon that teachers have to perform "on the fly". These interventions are maybe not sufficient, but all teachers will agree that they are necessary in language teaching. The problem is that, as Polguère (2004) has observed, right now, teachers are often times not trained enough to perform analysis of lexical phenomena quickly and accurately during class interactions. Lexicon is so large that there can't be a ready-to-use answer to every possible question teachers can be asked. Our objective is therefore to make teachers more autonomous by training them to observe and to analyze lexical phenomena so they can face any new lexical problem. We believe that the knowledge of the participants to our experiments, all experienced lexicographers, could benefit the teachers as much as other lexicographers. For the time being, when analyzing the data from experiments, we describe participant's every moves, without judging the efficiency or the relevance of each operation. Eventually, we'll have gathered enough data to derive models of approaches to solve different types of problems, or to accomplish different types of tasks. Teachers could study these models, and pedagogical activities could be created for them. In so doing, not only would they learn by example, they could also initiate themselves to lexicography. Of course, the goal here is not to turn teachers into lexicographers, but to give them some basic tools and a method to analyze lexical phenomena efficiently. Another implication of our work for pedagogy could aim at learners; some activities could be derived from the ontology to train them using dictionnaries, help them understand the strucutre of definitions, the concept of synonymy, etc.

Acknowledgement

We'd like to thank the Social Sciences and Humanities Research Council of Canada (SSHRC) for supporting this research, and the Département de linguistique et de traduction de l'Université de Montréal for a redaction bursary.

References

Atkins, Sue & Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford University Press, Oxford.

Béjoint, Henri. 2000. *Modern Lexicography: an Introduction*. Oxford University Press. Oxford.

Branch, Jennifer L. 2000. Investigating the Information-Seeking Processes of Adolescents: The Value of Using Think Alouds and Think Afters. *Library & Information Science Research*, 22 (4): 371-392.

De Pietro, Jean-François. 2003. L'enseignement du lexique en Suisse ou: comment en finir avec les listes à mémoriser? *Lettre de l'AIRDF*, 33: 12-18.

Ericsson, Anders K. & Herbert A. Simon. 1984. *Protocol Analysis: Verbal Reports As Data*. MIT Press. Cambridge. MA.

Ericsson, Anders K. & Herbert A. Simon. 1993. *Protocol Analysis: Verbal Reports As Data*. MIT Press. Cambridge. MA.

Fontenelle, Thierry (editor). 2008. *Practical Lexicography: A Reader*. Oxford University Press. Oxford.

Grossmann, F., Paveau, M.-A., Petit, G. 2005. *Didactique du lexique: langue, cognitions, discours*. ELLUG. Grenoble.

Landau, Sidney. 2001. *Dictionaries: The Art and Craft of Lexicography*. Cambridge University Press. Cambridge.

Mel'čuk, Igor. 2006. Explanatory Combinatorial Dictionary. In G. Sica, editor. *Open Problems in Linguistics and Lexicography*: 225–355. Polimetrica. Monza.

Mel'čuk, Igor, André Clas, & Alain Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*. Duculot. Paris/Louvain-la-Neuve. 1995.

Noy, Natalya F. & Deborah L. McGuinness. 2001. Ontology Development 101: A Guide to Creating Your First Ontology. *Stanford Knowledge Systems Laboratory Technical Report KSL-01-05* and *Stanford Medical Informatics Technical Report SMI-2001-0880*. Stanford University. Stanford. CA.

Polguère, Alain. 1994. Savoir consulter un dictionnaire, c'est bien...Savoir comment on le construit, c'est mieux. *Québec français*, 134: 68-70.

Russo, J. Edwards, Johnson, Eric J., & Stephens, Debra L. 1989. The Validity of Verbal Protocol. *Memory & Cognition*, 17 (6): 759-769.

Simard, Claude. 1994. Pour un enseignement plus systématique du lexique. *Québec français*, 92: 65-67.

Zgusta. Ladislav. 1971. *Manual of Lexicography*. Academia. Prague.

Meaning-Text-Theory and Lexical Frames

Bob Coyne
Columbia University
New York, NY, USA
coyne@cs.columbia.edu

Owen Rambow
Columbia University
New York, NY, USA
rambow@ccls.columbia.edu

Abstract

We discuss the relationship between FrameNet and Meaning-Text Theory. We propose that the notion of frame can be used in MTT in order to give the conversive lexical function both a broader and a more precise definition. We include a discussion of some issues related to importing the FrameNet lexicon to MTT.

1 Introduction

This paper proposes an alternate (or additional) structure for the lexicon. This alternative is directly inspired by FrameNet and results from an effort to relate the FrameNet (Baker et al., 1998) lexicon to Meaning-Text Theory (Mel'čuk, 1988) conceptualizations. The result is fully compatible with the spirit of MTT, we believe, but captures the key intuitions of FrameNet. These intuitions allow for two enhancements to the MTT approach to the lexicon: first, it provides for a finer-grained notion of "conversive" which can make more subtle semantic distinctions; second, it allows for a generalized notion of conversive lexical function which captures are a wider range of phenomena.

Specifically, this paper has four goals:

- We would like to introduce the MTT community to a theoretical approach to the lexicon and a practical resource (an English lexicon) which is complementary to work in the MTT framework.
- We make specific proposals about generalizing the CONV lexical function in order to make it better defined and more usable.
- We make a specific proposal about how the notion of "frame" can be integrated into the MTT approach.
- We describe an attempt at automatically converting a lexical resource to the MTT framework. However, the goal of this paper is theoretical, and we mention this work only to show how our analysis can allow us to relate lexical resources.

We concentrate on verbs in this paper, tough FrameNet includes all parts of speech, and the extension of our discussion to nouns and adjectives would be interesting. This is left to future work.

There has been some previous work in relating FrameNet to MTT (Alonso Ramos et al., 2008). This work differs in that Alonso Ramos et al. (2008) concentrate on paradigmatic lexical functions, while we concentrate on syntagmatic lexical functions. Our work is thus complementary to theirs.

This paper is structured as follows. In Section 2, we summarize FrameNet. We then show how frames can be integrated with the MTT lexicon in Section 3.1, and propose our generalized conversive lexical function in Section 3.2. We then describe some practical issues related to extracting lexical functions from the FrameNet lexicon in Section 4.

2 FrameNet

FrameNet (Baker et al., 1998) is a digital lexical resource for English that groups related words together into semantic frames. FrameNet currently contains over 10,000 lexical units, where a lexical unit is defined as a pairing of a word (noun, verb, adjective) with a sense or meaning. In addition, there can sometimes be more than one lexical unit per word sense, based on different perspectives of that shared meaning. For example, the same sense of the verb SHOOT is represented by three separate lexical units corresponding to *shoot the target*, *shooting the gun*, and *shoot the bullet*. Each lexical unit is contained in one of nearly 800 hierarchically-related semantic frames, where each frame represents shared meaning between the lexical units in that frame. In addition, each lexical unit contains a set of annotated sentences which map the sentences' constituent parts to their frame-based roles. FrameNet, in total, contains over 135,000 annotated sentences across all lexical units. Not all lexical units have been annotated. For example, of the approximately 4,100 verb lexical units in FrameNet, only about 2,800 have annotated sentences.

A FrameNet frame consists of a set of frame-based roles, called *frame elements* (FEs). For example, the COMMERCE_SELL frame includes frame elements for SELLER, GOODS, and BUYER. These and other FEs represent the key roles that characterize the meaning of the lexical units in that frame. Frames can contain any number of individual lexical units. The COMMERCE_SELL frame, for example, has lexical units for the words RETAIL, SELL, VEND, etc.

The exact expression of FEs for a given annotated sentence constitutes what FrameNet refers to as a *valence pattern*. In this paper we represent valence patterns as lists of FE and grammatical function (GF) pairs. Grammatical functions are subject (*ext*), object (*obj*), second object (*Dep/NP*), and various other dependent phrases (*Dep/to*, *Dep/on*, *Dep/with*, etc.) which designate the particular prepositional phrase type. [1] So, for the verb GIVE, the sentence *John gave the book to Mary* has the valence pattern of: ((Donor Ext) (Theme Obj) (Recipient Dep/to)). And *John gave Mary the book* has the valence pattern of ((Donor Ext) (Recipient Obj) (Theme Dep/NP)). Every verb typically has many valence patterns, representing the various ways that verb can be used in sentences.

FrameNet makes a distinction between "core" FEs (those that are unique or characteristic to the meaning of the frame) and "peripheral" frame elements (which do not uniquely characterize a frame). For example, TIME, LOCATION, and MANNER are typically peripheral FEs since they can be instantiated in any appropriate frame. In contrast, in the COMMERCE_BUY frame (which includes the verbs BUY and PURCHASE), the FEs for BUYER and GOODS are core since they are central and conceptually necessary to the meaning of that frame.

FrameNet frames are related to each other by a fixed set of frame relations. These allow us to find semantically related verbs across frames. In addition, since frames can give arbitrary names to their frame elements, frame relations are used to define the mapping between corresponding frame elements in the related frames. Some relevant frame relations are:

INHERITANCE: This relation represents an is-a relation between two frames. An example is the ESCAPING frame which inherits from the DEPARTING frame.

PERSPECTIVE_ON: This relation links perspectivized frames which represent two different points-of-view of some other neutral frame. For example the frames for verbs BUY and SELL are related as perspectives on the COMMERCIAL_TRANSACTION frame. Similarly, there are three frames for SHOOT corresponding to *shoot the target*, *shooting the gun*, and *shoot the bullet*. These three frames are related via the PERSPECTIVE_ON relation.

REFRAMING_MAPPING: It's not uncommon for lexical units to be moved into new frames. This frame relation remaps the names of FEs from one frame to another when frame boundaries have changed

[1] Note that the FrameNet syntactic annotation always calls the first NP following the verb the object, even in the double-object construction (*John gave Mary a book*), where the standard analysis would call the second NP the object and the first NP the indirect object. We follow the FrameNet annotation, despite the fact that the standard analysis is more appealing, because we need to work with the given annotation.

(Petruck et al., 2004). For example the FORGING frame (containing verbs such as FALSIFY, FAKE, and COUNTERFEIT) is related via the **REFRAMING_MAPPING** frame relation to the FEIGNING frame (which contains separate lexical units for some of the same verbs plus others for verbs such as AFFECT and PRETEND).

INCHOATIVE_OF and CAUSATIVE_OF: The relationship between stative frames and corresponding inchoative and causative frames is encoded with the **CAUSATIVE_OF** and **INCHOATIVE_OF** frame relations. For example, the verb COOL is represented by separate lexical units in the frames for CAUSE_TEMPERATURE_CHANGE (as in *John cooled the apple*) and INCHOATIVE_CHANGE_OF_TEMPERATURE (as in *The apple cooled quickly*). The frame CAUSE_TEMPERATURE_CHANGE is related by **CAUSATIVE_OF** to INCHOATIVE_CHANGE_OF_TEMPERATURE which in turn is related with **INCHOATIVE_OF** to the stative frame TEMPERATURE.

SUBFRAME: Some frames refer to sequences of other frames. These subframes are related to the parent frame via the SUBFRAME relation. For example, the frame CAUSE_IMPACT contains the lexical unit SLAM (as in *john slammed the car door*). This frame has a subframe IMPACT which contains a separate lexical unit for SLAM (as in *the door slammed shut*).

USING: The **USING** frame relation is used in cases in which a part of the child's meaning refers to the parent frame. For example, the COMMUNICATION_NOISE frame is used by verbs where communication takes place via a sound (e.g. the verb CLUCK in *"Sorry, Jimmy," the teacher clucks sympathetically at one unfortunate.*. To represent this dependency, it is related to the MAKE_NOISE frame via the **USING** frame relation.

SEE_ALSO: This frame relation is primarily intended for human users of FrameNet. But it sometimes encodes a direct relation between frames that would otherwise be only indirectly related. For example, the PERCEPTION_EXPERIENCE frame (e.g., verbs like SEE) and the PERCEPTION_ACTIVE frame (e.g. verbs like WATCH which imply a volitional act) are related indirectly via the **INHERITANCE** frame relation to their common parent frame PERCEPTION. They are also directly related by the **SEE_ALSO** frame relation.

Note that verbs in the same frame and related frames can vary significantly in meaning. For example, the SELF_MOTION frame contains a large number of verbs related only by the fact that the SELF_MOVER moves under its own power in a directed fashion without a vehicle. As a result, this frame contains strongly related verbs such as WALK and STROLL but also verbs with very different manner of motion such as SWIM and SWING.

By way of an example, we show how various frames related to the commercial transaction meaning are related in Figure 1.

3 Frames in MTT

In this section, we address the question how the notion of "frame" can be incorporated into MTT. In particular, we are interested in exploiting the relations between frames, and the extended notion of conversive that these relations allow.

3.1 Relating SemRs

In FrameNet, the core meaning of "frame" is provided by "semantic overlap". For atomic frames, the criteria for "semantic overlap" are very strict: two verbs can only be in the same atomic frame if they have the same number of actants, the same aspectual behavior, the same presuppositions, and if their (obligatory and non-obligatory) actants are interpreted in the same manner with respect to the underlying semantics. This latter point can be illustrated by purpose (or reason) clauses; this example is adapted from (Ruppenhofer et al., 2006) (the sign '#' denotes semantic infelicity):

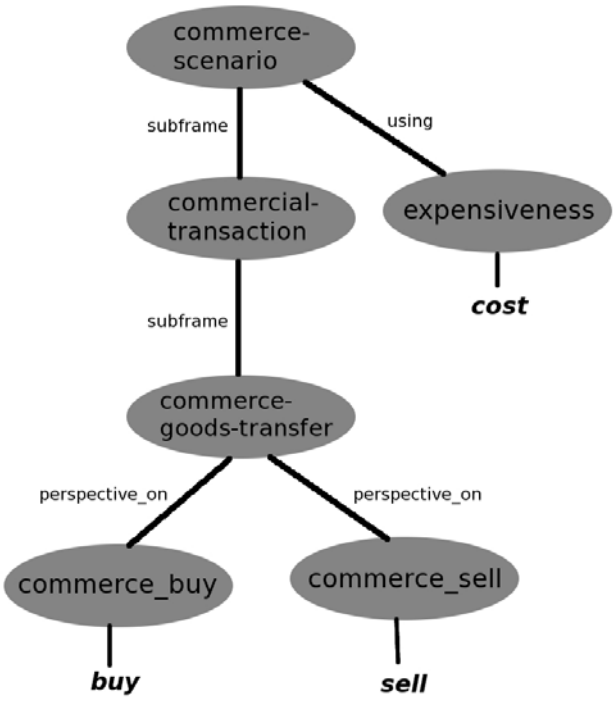


Figure 1: Relation between frames relating to commercial transactions

(1) a. Mary sold Peter an apartment because she wanted to make him happy/#he wanted to make her happy

b. Peter bought an apartment from Mary because he wanted to make her happy/#she wanted to make him happy

c. An apartment was sold to Peter by Mary/Peter was sold an apartment by Mary because she wanted to make him happy/#he wanted to make her happy

In (1a), the *because* clause can only refer to Mary's motivations,[2] and in (1b), the *because* clause can only refer to Peter's. As a result, as discussed by (Ruppenhofer et al., 2006), these two sentences are not paraphrases of one another.[3] Since BUY and SELL are conversives in MTT, this means that FrameNet atomic frames are more restrictive than MTT's notion of conversive in terms of semantic requirements. It also means that conversive do not always license paraphrases in MTT. We leave aside the issue of why the motivation clauses choose the (underlying) subjects, how MTT represents this, and why this fact is not relevant to the MTT definition of the basic conversive.

We define the atomic **lexical frame** to be the SemR L of a given lexeme, and we associate with it extensionally all lexemes that have exactly the same SemR L. We then relate lexical frames to each other using the following asymmetric specialization relations on semantic representations (SemRs), which we define in

[2]The variant marked with '#' are pathologically acceptable if contextually one may infer that Mary wanted to fulfill Peter's desire by letting him make her happy. However, it is still an explanation of Mary's action, not of Peter's,even though this commercial transaction involves actions by both Mary and Peter. Similar comments apply to the other sentences.

[3]1c) shows that this is not an effect of the *because* clause being oriented towards the surface subject – (1c) has the same interpretation as (1a), not as (1b).

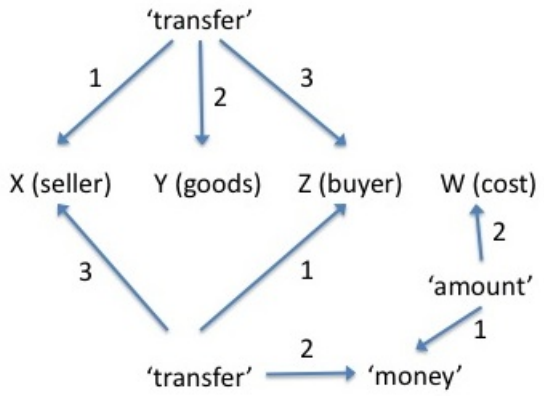


Figure 2: SemR for the COMMERCIAL_TRANSACTION frame

analogy to the frame relations of FrameNet (see Section 2). All these relations have in common that they relate a more specific SemR to a less specific SemR which contains less information (i.e., fewer nodes). They differ in what kinds of meanings are added, and thus in how the SemRs are related semantically. We define the extension of a less specific SemR to be the set of all lexemes associated with the frames it specializes. Note that we may encounter SemRs which themselves do not correspond to verbs, though usually the noun (or adjective-noun collocation) which is the name of a lexical frame is also a nominal lexeme with that meaning. (Recall, however, that we do not consider nouns in this paper.)

- **PERSPECTIVE_ON**: this relation links two SemRs which differ only in the fact that one records perspective, i.e. (typically), it attributes agentivity, while the other (less specialized) does not. It remains to be seen how this is expressed in MTT, this is not the topic of our paper. For example the frames for verbs BUY and SELL are both related as perspectives on the COMMERCIAL_TRANSACTION frame. The SemR of this frame is shown in Figure 2.

- **CAUSATIVE_OF**: this relation adds a causer to a SemR. For example, the verb COOL is contained in both CAUSE_TEMPERATURE_CHANGE (as in *John cooled the apple*) and INCHOATIVE_CHANGE_OF_TEMPERATURE (as in *The apple cooled quickly*). The frame CAUSE_TEMPERATURE_CHANGE is related by **CAUSATIVE_OF** to INCHOATIVE_CHANGE_OF_TEMPERATURE. Determining whether an actant is a causer is not always straightforward, see (Iordanskaja and Mel'čuk, 2002) for a discussion.

- **USING**: The **USING** frame relation is used when a SemR specializes another SemR by adding a completely new purpose or function. For example, the COMMUNICATION_NOISE frame is used by verbs where communication takes place via a sound, and it is related to the MAKE_NOISE frame via the **USING** frame relation.

- **SUBFRAME**: Some SemRs describe sequences of events, one of which is is described by another SemR. These subframes are related to the parent frame via the SUBFRAME relation. For example, the frame CAUSE_IMPACT contains the lexical unit SLAM (as in *John slammed the car door*). This frame is a subframe of IMPACT which contains a separate lexical unit for SLAM (as in *the door slammed shut*).

- **INHERITANCE**: This relation represents a generic is-a relation between two SemRs and is used when one of the more specific ones does not apply. An example is the SemR for ESCAPING which inherits

from the SemR for DEPARTING, adding the meaning that the departure required circumventing barriers aimed at preventing it.

3.2 Extending the CONV Lexical Function

The conversive lexical function preserves meaning but changes the syntactic realization of arguments. Given our fine-grained typology of relations between SemRs sketched above, we can be more precise about what sort of semantic relation exists between two lexemes related by a conversive lexical function. Recall that we have defined a lexical frame as a SemR and the set of verbs whose SemRs are related to this SemR in one of the specializations or generalizations listed in Section 3.1. Now, as the SemRs of two verbs start to differ more and more, the conversive becomes less and less meaning-preserving. We propose to indicate this by including, as a superscript on the CONV function name, the list of specializations that the SemRs undergo in order to become the SemR of the same frame.

Consider the classic pair of BUY andSELL. As can be seen from Figure 1, these two verbs are not in the same atomic frame, but they are both in the COMMERCE_GOODS_TRANSFER frame. Their atomic frames are both related to their shared frame through the **PERSPECTIVE_ON** relation. We thus get:

X buys Y from Z for W (=cost)
Z sells Y to X for W
$\text{CONV}_{3214}^{\text{persp}}$(BUY) = SELL

The first actant of BUY corresponds to the third actant of SELL, and the third actant of BUY corresponds to the first actant of SELL. The second and fourth actants correspond to the second and fourth actants, respectively. The superscript "persp" tells us that the two lexemes are identical in meaning, subject to the specified argument permutation, except for the agentive perspective, and can be substituted in paraphrases except when this agentivity comes into play (say, in the presence of purpose clauses).

We now must address a technical limitation of the conversive lexical function. In its presentation to date, it relates only verbs which have the same number of actants: the subscript indicates the permutation of the actants, using standard mathematical notation for permutations. (A permutation is a total bijective function from a set onto itself.) For example, BUY and SELL both have four syntactic actants; but COST has at most three actants (*the goods cost the buyer the price*). Thus, their actants cannot be related by a permutation, and the verbs cannot be related by a CONV lexical function. This seems an artificially imposed restriction: the two verbs share some common meaning in their SemRs. Note that for COST, the missing actant is in fact implied: there can be no cost situation without a seller. It is just not realizable syntactically using COST as the main verb. We therefore propose a second extension to the CONV lexical function notation, which allows for any mapping between the set of actants. We introduce the following, very explicit notation:

$\text{CONV}_{1\to\emptyset,2\to1,3\to3,4\to2}^{\text{persp,subframe,using}}$ (SELL) = COST

The subscript tells us exactly and explicitly how each actant of SELL is mapped to an actant of COST; actants which COST does not realize are given null targets. Here, the superscripts tell us that the meaning has changed in several respects: there has been a change in agentivity (in fact, COST has no agentive actant); there has been an elimination of several parts of the meaning which correspond to substeps (the transfer of the bought goods is not part of the meaning of COST); and a core meaning is used for a different purpose.

It is clear that, with our new generalized conversive, we can easily express the standard permutation-based conversives as well:

$\text{CONV}_{1\to3,2\to2,3\to1,4\to4}^{\text{persp}}$ (SELL) = BUY

However, in addition, our new extended conversive lexical function notation handles several cases which the permutation-based one does not(we omit the superscripts as they are not relevant to this discussion).

1. Many verbs have optional arguments: for example, for BUY, arguments 3 and 4 are each individually optional. In the MTT dictionary, this is indicated in the valency pattern which relates the deep-syntactic representation to the surface-syntactic representation. While this is perfectly good from a theoretical perspective, it may be useful to consolidate the knowledge about possible argument realizations in a single part of the lexicon: if our extended conversive can relate BUY to COST (which has fewer actants), it may also be expedient to relate BUY with four actants to BUY with two or three actants. Of course, this is still the same lexeme, it is just a new technical means of indicating that an actant is optional. (The "inher" superscript tells us that some meaning is removed.)

   $\text{CONV}^{\text{inher}}_{1\rightarrow1,2\rightarrow2,3\rightarrow\emptyset,4\rightarrow4}$ (BUY) = BUY
   $\text{CONV}^{\text{inher}}_{1\rightarrow1,2\rightarrow2,3\rightarrow3,4\rightarrow\emptyset}$ (BUY) = BUY

2. Diathesis alternations of the type studied by Levin (1993) are a complex issue representationally. In these cases, semantic arguments can appear in different deep-syntactic argument positions. For example, in English we have *John loaded the truck with hay* as well as *John loaded hay into the truck*.[4] One approach is to have one lexeme but two valency patterns. The relationship between these two valency patterns can be expressed with an extended conversive, if we consider there to be underlyingly four actant positions such that the second actant is not mapped to any semantic argument, only the first, third, and fourth.

   $\text{CONV}^{\text{atomic}}_{1\rightarrow1,2\rightarrow3,3\rightarrow\emptyset,4\rightarrow2}$ (LOAD) = LOAD
   $\text{CONV}^{\text{atomic}}_{1\rightarrow1,2\rightarrow4,3\rightarrow2,4\rightarrow\emptyset}$ (LOAD) = LOAD

   As in the case of optional arguments, the knowledge about which semantic arguments can be expressed how is concentrated in one place, namely in the extended conversives. The "atomic" superscript tells us that the SemRs are identical.

Thus, in summary, we have proposed an extended notation for lexical functions, which is based on the notion of lexical frame, and which is both more flexible as to the realization of arguments,and more precise as to the nature of the meaning shared by the two related lexemes.

4 Converting the FrameNet Lexicon to MTT Relations

To generate paraphrase transformations for a given verb we first determine its FrameNet frame and then find related frames using the following frame relations: Null (i.e. look at other verbs within the same frame); **PERSPECTIVE_ON**; **USING**; **REFRAMING_MAPPING**; **INCHOATIVE_OF**; **CAUSATIVE_OF**; **INHERITANCE**; **SUBFRAME**; and **SEE_ALSO**. We then collect a list of potential target verbs in those related frames. Note that the target verb can be the same as the source verb. This allows us produce paraphrases involving diathesis alternations and argument ommission.

In selecting potential verbs we rely upon WordNet to test for different types of synonymy. WordNet (Fellbaum, 1997) is a lexical database for the English language. It groups nouns, verbs, and adjectives into sets of synonyms called synsets and arranges those synsets into a hypernym/hyponym hierarchy. So, for example, the synonyms SHOPKEEPER and STOREKEEPER are in the same synset. They are hyponyms of MERCHANT and hypernyms of FLORIST and TOBACCONIST. Note that WordNet synsets and FrameNet lexical units for a given lexeme don't usually correspond exactly. For example, WordNet has a single synset for the verb CLANG, while FrameNet has four lexical units representing separate frames for CAUSE_TO_MAKE_NOISE, CAUSE_IMPACT, IMPACT, MAKE_NOISE, and MOTION_NOISE.

In order to ensure that verb pairs are extended conversives, target verbs must satisfy one of the following conditions:

[4] We leave aside issues of restrictions on argument realizations in different patterns, for example on pronouns.

- The source and target verbs are in the same WordNet synset and hence treated as synonymous. This allows verbs such as HALT and STOP to be paired.

- The source and target verbs are respectively WordNet hyponyms or hypernyms of each other. So, for example, GO is a hypernym of WALK. Thus, sentences like *John walked to the store* and *John went to the store* could be used to describe the same event.

- The source and target verbs are related via the **PERSPECTIVE_ON** frame relation. This covers the case where the two verbs have different surface meaning but denote the same underlying event from different points of view. For other frame relations we require one of the synonym, hypernym, or hyponym relations as described above.

Note that in this strategy, we will miss some valid paraphrase pairs because of the granularity of meaning in WordNet. For example, in WordNet, BATHE and WASH are not synonyms, hypernyms, or hyponyms of each other. Instead they are both children of CLEANSE which also has unrelated children such FLOSS. As a result we would require some other technique or resource to identify close siblings such as BATHE and WASH in order to include them in our paraphrase pairs.

After collecting the possible target verbs, we then identify all valence patterns for the given source verb. These represent the possible left-hand sides of the transformations. Then, for each of these left-hand side patterns, we collect the right-hand sides from target verb list which have compatible valence patterns. Since we are primarily concerned with paraphrase, the right-hand side valence patterns must only reference FEs that are explicitly expressed in the given left-hand side.

Constructing the valence pattern from the annotation is straightforward, as both the grammatical function and the FE are marked, except for one very important aspect: grammatical voice. The active/passive alternation can be seen as an entirely productive verb alternation in English, and it would make no sense to suggest that every valence pattern for every verb has two additional variants (the passive, and the passive with *by*-agent). Instead, we want to normalize for voice, i.e., we want to always represent the valence pattern for active voice. This is non-trivial, because the syntactic annotation of FrameNet does not include a feature for voice, and the provided grammatical function annotation is for the surface grammatical function. We have implemented a series of heuristics that exploits the part-of-speech and grammatical function annotations, as well as the annotation for missing arguments in passives without *by* agents. However, some cases are impossible to disambiguate for a variety of reasons, including a fair number of examples in which the main verb form is not disambiguated between past tense and past participle and there is no auxiliary (reduced passive relative clause or conjunctions). Thus, grammatical voice is the major source of errors for us in determining the valence pattern.

For example, the verb GIVE has the following left-hand side patterns (each pattern consisting of FE and GF pairs):

Left-hand side pattern	**Example sentence**
((Donor Subj) (Recipient Obj) (Theme Dep/NP))	*John gave Mary the book*
((Donor Subj) (Theme Obj) (Recipient Dep/to))	*John gave the book to Mary*
((Donor Subj) (Theme Dep/of) (Recipient Dep/to))	*John gave of his time to people like Mary*
((Donor Subj) (Recipient Dep(to)))	*John gave to the church*

Verbs in related frames have the following valence patterns. As a result, the possible right-hand side patterns will be drawn from the following list:

Possible right-hand side patterns	
((Donor Subj) (Theme Obj) (Recipient Dep/to))	((Donor Subj) (Theme Obj))
((Donor Subj) (Recipient Obj) (Theme Dep/NP))	((theme Subj))
((theme Subj) (Recipient Dep/to))	((Donor Subj) (Recipient Obj) (Theme Dep/to))
((Donor Subj) (Theme Obj) (Recipient Dep/on))	((Donor Subj) (Recipient Obj))
((Donor Subj) (Recipient Obj) (Theme Dep/with))	((Donor Subj) (Theme Obj) (Recipient Dep/for))
((Donor Subj) (Theme Obj) (Recipient Obj))	((Donor Subj) (Theme Obj) (Recipient Dep/Poss))
((Donor Subj) (Recipient Dep/to))	((Donor Subj) (Theme Dep/of) (Recipient Dep/to))
((Recipient Subj) (Theme Obj))	((Donor Subj) (Recipient Obj) (Theme Dep/VPto))
((Donor Subj) (Theme Obj) (Recipient Dep/upon))	((Donor Subj) (Recipient Obj) (Theme Dep/the))

For each possible valence pattern on the left-hand side we collect all valence patterns for the right-hand side that contain only FEs present in the given left-hand side valence pattern. For each of these target valence patterns we list all corresponding verbs (including their lexical unit ID and their frame) along with a matching annotated sentence. In presenting these examples of extended conversives we use the actual annotated sentences associated with the given lexical units in the FrameNet corpus. So while the overall pairs of sentences don't have the same meaning, they contain verbs with shared meanings, and all examples are naturally ocurring. For example, here's a single output entry for the verb GIVE:

FROM	
Pattern::	((Donor Subj) (Recipient Obj) (Theme Dep/NP))
Verb:	[LU-4344 "give.v" Giving]
Example:	*Katy and Jamie got ready very quickly and Mum* **gave** *each of them two wee spoons.*
TO	
Pattern:	((Donor Subj) (Theme Obj) (Recipient Dep/to))
Verb:	[LU-4344 "give.v" Giving]
Example:	*They wrapped it up and* **gave** *it to her , and it did have a head like a baby.*
Verb:	[LU-5344 "donate.v" Giving]
Example:	*Ralph and Philip are looking for local sponsors to* **donate** *money to their twin charities.*

In constructing these valence patterns we only consider core FEs since these will be characteristic of the verbs in question. Also, in collecting the valence patterns in the mappings from one frame to another, the FE names will often be different. For example, the COMMERCE_SELL frame (used by the verb SELL) has a FE called SELLER, but in the Expensiveness frame (used by the verb COST as in *the book costs 10 dollars*) this identical role is called PAYER. FrameNet's frame relations specify how these different FE names get mapped into each other. We use this information to automatically normalize the names to the namespace of the parent frame. It is the normalized FE names that are output in the paraphrase transformation patterns.

By relaxing the constraint that right-hand side valence patterns can only reference FEs explicitly expressed on the left-hand side, we get alternations such as the following:

FROM	
Pattern	((Agent Subj) (Projectile Obj) (Firearm Dep(from)))
Verb	[LU-5314 "shoot.v" Shoot_projectiles]
Example	*You can use it to* **shoot** *heavy balls of metal from large guns.*
TO	
Pattern	((Agent Subj) (Firearm Obj))
Verb	[LU-5315 "shoot.v" Use_firearm]
Example	*Alex Household must have said it just before he* **shot** *the gun;*

5 Conclusion and Future Work

This paper has discussed the relation between FrameNet and MTT, specifically the organization of the lexicon proposed by the two theories. We have suggested that FrameNet;s hierarchical notion of meaning and its fine-grained inventory of relations between meanings can be expressed in MTT as well, leading to an extended notion of the conversive lexical function which is both more inclusive and more precise. Future work will require a more detailed elaboration of the SemRs of FrameNet frames; in this manner, MTT will contribute a clearer (and computationally interesting) definition of lexical meaning for FrameNet.

Acknowledgments

We would like to thank Sylvain Kahane for useful feedback on the ideas presented in this paper, and two anonymous reviewers for their useful comments. This work was supported in part by the Defense Advanced Research Projects Agency (DARPA) under contract number HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of DARPA.

References

Alonso Ramos, Margarita, Owen Rambow, and Leo Wanner. 2008. Using semantically annotated corpora to build collocation resources. In *Proceedings of LREC*, Marrakesh,Morocco.

Baker, Collin F., J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 86–90, Montréal.

Fellbaum, Christiane. 1997. *WordNet: An Electronic Lexical Database*. MITPress, Cambridge, MA.

Iordanskaja, Lidija and Igor Mel'čuk. 2002. Conversif ou causatif? *Cahiers de Lexicologie*, 80(1):105–119.

Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. The University of Chicago Press.

Mel'čuk, Igor A. 1988. *Dependency Syntax: Theory and Practice*. State University of New York Press, New York.

Petruck, M.R.L., C. Fillmore, C. Baker, M. Ellsworth, and J. Ruppenhofer. 2004. Reframing framenet data. In *Proceedings of The 11th EURALEX International Congress*, pages 405–416, Lorient, France.

Ruppenhofer, Josef, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2006. Framenet ii: Extended theory and practice. Technical report, ICSI. http://framenet.icsi.berkeley.edu/book/book.html.

On Speaker's Stance Meaning of Discourse

Alexander Dikovsky
LINA CNRS UMR 6241, Université de Nantes
alexandre.dikovsky@univ-nantes.fr

Abstract

We outline a semantics for discourse from the speaker's stance. It's expressions, Discourse Plans, explicitly mark co-reference and present the verbal's predication in diverse diatheses. From the discourse plans, this semantics computes a relational structure representing verbals by unique canonical relations and interpreting nominals through their set extensions, taking into account plurality.

1 Introduction

Let us see the following text. *Currently, insurers can increase premiums by (levying surcharges if they determine (a driver)$_{\Downarrow_x}$ is more than 50 percent to blame for a collision)$_{\Downarrow_e}$. (Such penalties)$_{\Downarrow_p\ (e \in p)}$ often cost* $0_{\Uparrow_x}$ *hundreds of dollars annually for up to six years. (About half of (the 50,000 cases disputed each year)$_{\Downarrow_c\ (c \sim p)}$)$_{\Downarrow_{c_h}\ \mathbf{part}^{\sim}_{0.5}(c_h, c)}$ are overturned by the appeals board. (Those drivers)$_{\Downarrow_d\ \mathbf{of-concern}(d, c_h)}$ are issued refunds*. [The Boston Globe, March 2, 2009].

Here are tagged the constituents describing entities and events related between them within this discourse. Suppose that $\Downarrow_x$ in *(a driver)*$_{\Downarrow_x}$ means something like: "a new semantical object x will identify the entity denoted by the selected occurrence of *a driver* in the discourse" and that $\Uparrow_x$ is the object identified by x. Then e identifies the *levying* event, which is a kind of the *penalties* p that cost much to the *drivers* x (elided in the text). Further, c are the *disputed cases*, $c \sim p$ means that c identifies the same object as p and c_h identifies *about half of them..overturned...* Finally, d are the *drivers* concerned with the cases c_h.

One can see that this tagging goes beyond the anaphora. Where does it come from? In contrast to the logical semantics of discourse, such as DRT (Kamp et al.,), it is not supposed to be *computed from the discourse*. On the contrary, we proceed from the assumption that *this tagging is given*: it represents elements of a *speaker's discourse plan* from which the discourse is to be *realized*. This is one of the roles to be played by a semantic representation of discourse in the context of the Meaning-Text Theory. In this role, the representation serves as a semantical notation.[1] But it should also provide relational structures, let us call them *contexts*, evolving in the discourse and suited for logical analysis. Only meaning representations playing these two roles may pretend to represent the *speaker's stance meaning of discourse*. Our example shows a specificity of the contexts. On the one hand, entities are treated as *sets evolving in the discourse*. On the other hand, *events may also behave as entities*, for instance, become elements of other entities-sets. The other specificity is less evident. It is implied, using Occam's Razor, by the speaker's stance itself. In contrast with the hearer's stance, the speaker's one needs not a reference analysis (the speaker disposes of complete knowledge of reference to express). Context consistency is also not required: the facts are postulated.[2]

[1] In (Dikovsky, 2007) is studied a formal system representing MTT in terms of finite tree transducers on discourse plans.

[2] This doesn't prevent from inclusion of the (extra-linguistic) consistency check in an implementation.

The speaker's stance semantics outlined in this paper is *object oriented* in the sense that entities and events are uniquely identified by invariable semantical objects characterized by values of *attributes* and by an *extension* which is an evolving set. In fact, we outline two semantics. The first is *static*: it defines objects' extensions in a fixed context. The other semantics is *dynamic*. It defines both, the evolution of the contexts, and the objects' extensions. This semantics is only outlined because of space limits and will be published elsewhere. The two semantics prove to be equivalent in every context.

2 Discourse Plans

Expressions of our semantics, called *discourse plans* (DP), are already published (see (Dikovsky, 2003; Dikovsky and Smilga, 2005; Dikovsky, 2007)). Below, a *discourse* is seen as a sequence of DP. Here we show and comment main features of DP using an example of a discourse consisting of DP of two sentences among which the first is a variant of "donkey sentences" borrowed from (Kamp et al.,) (see Fig. 1,2).

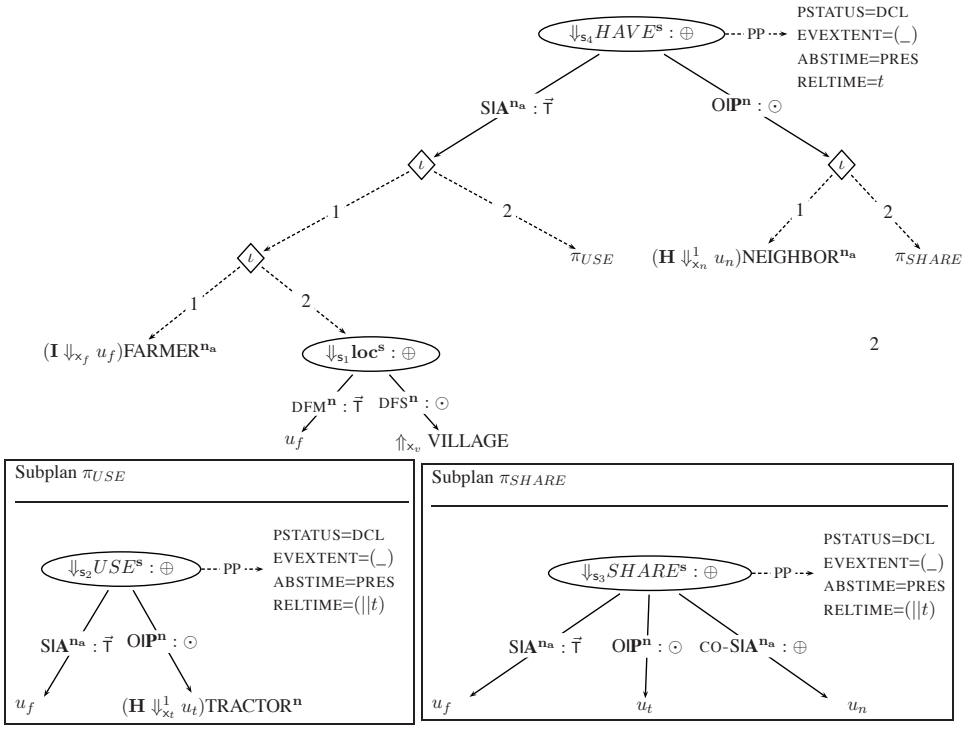


Figure 1. A DP of *Every farmer in the village, who uses a tractor, has a neighbor with whom he shares it.*

In DP, all semantemes have lexical types. Primitive types are partially ordered by a genericity order $\preceq$ ($\mathbf{u} \preceq \mathbf{v}$ means $\mathbf{u}$ is a case of $\mathbf{v}$). In composite types $(\phi \to \mathbf{v})$, argument types are identified by sorts.

Example 1 Some nominal types and nominals. $\mathbf{n}$ *(*nominals*),* $\mathbf{n_a} \preceq \mathbf{n}$ *(*animated nominals*),* $\mathbf{n_{count}} \preceq \mathbf{n}$ *(*countable nominals*),* $\mathbf{n_{ncount}} \preceq \mathbf{n}$ *(*uncountable nominals*) are examples of nominal types.* $\mathbf{nct} = (\text{STATE}^{\mathbf{a_{grad}}} \to \mathbf{n_{ncount}})$ *(cf.* hot milk*),* $\mathbf{ves} = (\text{CONTENTS}^{\mathbf{n_{ncount}}}; \text{FULLNESS}^{\mathbf{a_{grad}}}\text{QUANT}^{\mathbf{a_{card}}} \to \mathbf{n_{vessel}})$ *(cf.* two full glasses of beer*) are compound nominal types (*CONTENTS *is its core argument).* $MILK^{\mathbf{nct}}$, $SAND^{\mathbf{nct}}$ *are nominals of type* $\mathbf{nct}$. $GLASS^{\mathbf{ves}}$, $PACK^{\mathbf{ves}}$ *are nominals of type* $\mathbf{ves}$.

Some attributor types: $\mathbf{a}$ *(*attributors*),* $\mathbf{a_{grad}} \preceq \mathbf{a}$ *(*gradable attributors*, cf.* $RED^{\mathbf{a_{grad}}}$, $FAST^{\mathbf{a_{grad}}}$*),* $\mathbf{a_{degr}} \preceq \mathbf{a}$ *(*degree attributors*, cf.* $VERY^{\mathbf{a_{degr}}}$, $A\_BIT^{\mathbf{a_{degr}}}$*),* $\mathbf{a_{ord}} \preceq \mathbf{a}$ *(*ordinal attributors*, cf.* $FIRST^{\mathbf{a_{ord}}}$*)* $\mathbf{a_{card}} \preceq \mathbf{a}$ *(*cardinal attributors*, cf.* $FIVE^{\mathbf{a_{card}}}$, $MANY^{\mathbf{a_{card}}}$*),* $\mathbf{a_{prec}} \preceq \mathbf{a}$ *(*precision attributors*, cf.* $ABOUT^{\mathbf{a_{prec}}}$, $NEARLY^{\mathbf{a_{prec}}}$*).*

Verbals have types $(\phi \to \mathbf{s}')$, where $\mathbf{s}' \preceq \mathbf{s}$ and $\mathbf{s}$ is the *sentential type*. Their argument structure is determined by diatheses and diathetic shifts. For that, the sorts are divided into *roles* R and *attributes* A. We use the most generic roles such as Sl**A**(*subject-agent*), Ol**P**(*object-patient*), etc. The roles identify the *core* arguments, the attributes identify circumstantials and *propositional parameters* (PP). In Fig. 1-3 DP are presented in a graphical form where solid lines labeled with roles link verbals to their core arguments and

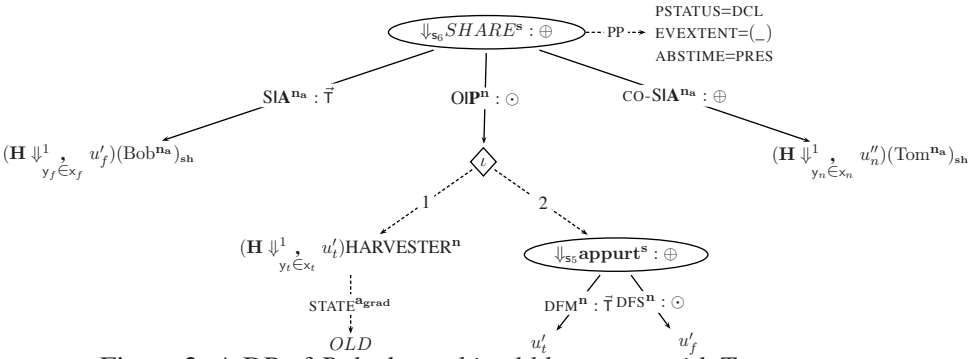


Figure 2. A DP of *Bob shares his old harvester with Tom.*

dashed lines labeled with attributes link semantemes to their circumstantials/qualifiers (e.g., $OLD^{\mathbf{a_{grad}}}$ represents the value of attribute STATE of HARVESTER in Fig. 2). The DP in Fig. 1,2 use several PP-attributes: PSTATUS (declarative in Fig. 1,2), SIGN (*positive*/*negative*), EVEXTENT, a generalized aspect (*continuous interval* in Fig. 1,2), time parameters ABSTIME (e.g., PRES) and RELTIME (*relative time*) etc.

The verbal SHARE has in Fig. 1,2 the type $((\text{SIA})^{\mathbf{n_a}}(\text{OIP})^{\mathbf{n}}(\text{CO-SIA})^{\mathbf{n_a}} \rightarrow \mathbf{s})$. If a verbal V has several types: $types(V) = \{\mathbf{t_0}, \ldots, \mathbf{t_p}\}$, we call them *diatheses* of V. One of the diatheses, $\mathbf{t_0}$, is selected as *canonical*. E.g., the canonical diathesis of the verbal OPEN is $((\text{SIA})^{\mathbf{n_a}}(\text{OIP})^{\mathbf{n}}(\text{INS})^{\mathbf{n}} \rightarrow \mathbf{s_{eff}})$ (cf. *John opened the door with the key*), but in *The key opened the door* it has an object alternation diathesis $\mathbf{dalt} = ((\text{SIA})^{\mathbf{n}}(\text{OIP})^{\mathbf{n}} \rightarrow \mathbf{s_{eff}})$. Non canonical diatheses are the result of transformations of $\mathbf{t_0}$, called *diathetic shifts*. The diathetic shifts correspond to core arguments' alternations/elimination caused at surface by change of mode, nominalization, conversion to infinitive, etc.

Definition 1 *Let* $\mathbf{t_0} = (\text{R}_1^{\mathbf{u_1}} \ldots \text{R}_n^{\mathbf{u_n}}; \text{A}_1^{\mathbf{v_1}} \ldots \text{A}_m^{\mathbf{v_m}} \rightarrow \mathbf{v})$ *be the canonical diathesis of* V *and* $\mathbf{t_i} = ((\text{R}')_1^{\mathbf{u'_1}} \ldots (\text{R}')_k^{\mathbf{u'_k}}; \text{A}_1^{\mathbf{v_1}} \ldots \text{A}_m^{\mathbf{v_m}} \rightarrow \mathbf{v'})$ *be some other its diathesis. Then* $D_i = (\mathbf{t_i}, \mathbf{d}_i)$ *is a* diathetic shift *of* $\mathbf{t_0}$ *if* $\mathbf{d}_i : \{1, \ldots, k\} \overset{1-1}{\rightarrow} \{i_1 \ldots, i_k\}$, *for* $1 \leq i_1 < \ldots < i_k \leq n$, *is a bijection preserving types:* $\mathbf{u'_j} = \mathbf{u_{i_j}}$, $1 \leq j \leq k$. *We call this bijection* argument shift *and denote it by* $\mathbf{d}_i : k \overset{1-1}{\mapsto} n$. *The non-canonical diathesis* $V^{\mathbf{t_i}}$ *resulting from* $V^{\mathbf{t_0}}$ *through diathetic shift* D_i *is denoted* $V[\mathbf{d}_i]^{\mathbf{t_i}}$ *and called a* derivative *of* $V^{\mathbf{t_0}}$.

Example 2 *For the verbal OPEN, the argument shift* $\{1 \mapsto 2, 2 \mapsto 1, 3 \mapsto 3\}$ *transforms its canonical diathesis into the diathesis of passive* $((\text{SBJ})^{\mathbf{n}}(\text{AGT})^{\mathbf{n_a}}(\text{INS})^{\mathbf{n}} \rightarrow \mathbf{s_{eff}})$ *as in the sentence* The door was opened with the key by John's girl-friend *and* $\{3 \mapsto 1, 2 \mapsto 2\}$ *transforms it into the diathesis of alternation* $\mathbf{dalt}$ *shown above.*

In DP, the diathetic shifts are represented as assignments to core arguments of new roles and of *communicative ranks* ($\vec{\text{T}}$: *topic*, $\odot$: *focus*, $\oplus$: *background*). Assignment of rank $\ominus$ (*periphery*) to an argument causes its elimination. E.g. SHARE has in Fig. 1,2 its canonical diathesis $((\text{SIA})^{\mathbf{n_a}}(\text{OIP})^{\mathbf{n}}(\text{CO-SIA})^{\mathbf{n_a}} \rightarrow \mathbf{s})$ for which the canonical rank assignment is SIA $\hookleftarrow \vec{\text{T}}$, OIP $\hookleftarrow \odot$, CO-SIA $\hookleftarrow \oplus$. A different role/rank assignment $\emptyset \hookleftarrow \text{SIA}_{\ominus}$, SIA $\hookleftarrow \text{OIP}_{\vec{\text{T}}}$, CO-SIA $\hookleftarrow$ CO-SIA$_{\odot}$ would transform the canonical diathesis into a diathesis of passive: $((\text{SIA})^{\mathbf{n}}\ (\text{CO-SIA})^{\mathbf{n_a}} \rightarrow \mathbf{s})$ (SIA-argument is eliminated by assignment of $\ominus$; assignment of $\vec{\text{T}}$ to the OIP-argument promotes it to the SIA-position). This assignment corresponds to the *argument shift* $\{2 \mapsto 1, 3 \mapsto 2\}$.

The intuitive reading of operator ι is "*such object .. that ..*".

Expressions $\mathbf{D}_{\mathsf{x}_f} = (\mathbf{H} \Downarrow_{\mathsf{x}_f} u_f)$ in $\mathbf{D}_{\mathsf{x}_f}$FARMER and $\mathbf{D}_{\mathsf{y}_f} = (\mathbf{H} \Downarrow^1_{\mathsf{y}_f \dot{\in} \mathsf{x}_f} u'_f)$ in $\mathbf{D}_{\mathsf{y}_f}(\text{Bob}^{\mathbf{n_a}})_{\mathbf{sh}}$ are *determiners* creating and relating objects. Created objects may be referred by other determiners in subsequent discourse. E.g. the constructor $\Downarrow^1_{\mathsf{y}_f \dot{\in} \mathsf{x}_f}$ "creates" a new object for the shifter name $(\text{Bob}^{\mathbf{n_a}})_{\mathbf{sh}}$ and "refers" to the object which binds x_f. Constructor $\Uparrow_{\mathsf{x}}$ just refers to x. E.g., $\Uparrow_{\mathsf{x}_v}$ VILLAGE in Fig. 1 gives the object o_v previously created by $\Downarrow_{\mathsf{x}_v}$ VILLAGE. Constructors $\mathbf{H}$ and $\mathbf{I}$ used in determiners define the way the referred objects are accessed. The former (*holistic*) provides access to the object itself, whereas the latter (*individual*) to the object's extension elements. The full DP syntax may be seen in the definition of the static semantics.

3 Fundamentals of DP Semantics

General notions. We define two DP semantics: one *dynamic*, the other *static*. Both are defined in a subset of the set theory extended with specific constants: $R_{\mathbf{g}}$ (*global object references*), $R_{\mathbf{l}}$ (*local object references*), $\mathbf{O^t}$ (*object identities* of type $\mathbf{t}$, $\mathbf{O}$ is the union of all $\mathbf{O^t}$), *lexical class constants* in $LC = \{L_W \mid W \text{ is a semanteme}\}$, $\perp \notin \mathbf{O}$ (an *"uncertain value"*).

As show the examples in Fig. 1,2, DP do not use quantifiers and object variables. Instead they use determiners $\mathbf{D}_{\mathsf{x}}$, where x is a global reference in $R_{\mathbf{g}}$. The dynamic semantics is relativized to *dynamic contexts* (*d-contexts*). When the semantics of a DP $\pi = \mathbf{D}_{\mathsf{x}}\pi'$ is computed in a d-context Σ, it assigns to π a new object $o \in \mathbf{O}$ (a *realization* of π in the discourse), changes Σ to a new d-context Σ', binds x with o and assigns to o a set value $|o|^{\Sigma'}$, its *dynamic extension* (*d-extension*). In other words, the effect of the DP π in Σ may be seen as a *transition* $(\pi)_{\Sigma}^{\Sigma'}$ from Σ to Σ'. The d-extension of π is relativized to the initial d-context: $|\pi|_{\Sigma}^{\Sigma'} =_{df} |o|^{\Sigma'}$. The static semantics is insensitive to context transitions. It applies to a DP π in a *static context* (*s-context*) σ where it inductively computes a set value $\|\pi\|^{\sigma}$ called *static extension* (*s-extension*) of π. The two semantics are related through a tight correspondence between d- and s-contexts.

Definition 2 *A* d-context *is a finite structure $\Sigma = (D, I)$, where D is a finite collection of sets and I is a finite function from constants to sets in D with four particular restrictions: $\gamma_{\Sigma} = I\restriction R_{\mathbf{g}}$ (*global assignment*), $\lambda_{\Sigma} = I\restriction R_{\mathbf{l}}$ (*local assignment*), $\theta_{\Sigma} = I\restriction \mathbf{O^n}$ (*nominal objects' evaluation*) and $\hbar_{\Sigma} = I\restriction LC$ (*horizon line *of* Σ*). The finite structure $\sigma =< \gamma_{\Sigma}, \lambda_{\Sigma}, \theta_{\Sigma}, \hbar_{\Sigma} >$ is the* s-context corresponding to Σ.

$\gamma_{\Sigma}(\mathsf{x}) = o$ means that the global reference x is bound with the object o, $\lambda_{\Sigma}(u) = s$ means that the local reference u is bound with the set s, $\theta_{\Sigma}(o) = s$ means that the nominal type object o has d-extension $|o|^{\Sigma} = s$ and $\hbar_{\Sigma}(L_W) = s$ means that s is the part of the d-extension of lexical class L_W "accessible" in Σ. So the corresponding d-context and s-context share the four functions. Context transitions in the dynamic semantics correspond to updates of some of them.

Elements of lexical semantics. DP semantics rests upon a set of lexical axioms. The axioms introduce lexical class constants L_W for semantemes $W^{\mathbf{t}}$ of type $\mathbf{t}$ and relates with them a set of functions (*attributes*). For space reasons, we only cite some their consequences necessary to understand semantical definitions.

We suppose that every semanteme W has a unique set code $W^{\star}$. Let $LEX(\mathbf{u})$ denote the set of all DP semantemes of types $(\phi \rightarrow \mathbf{u})$ or $\mathbf{u}$. First of all, we suppose that the lexical classes representing types consist of objects of these types: $L_{\mathbf{u}} \subseteq \mathbf{O^u}$. Attributor type objects are particular: every attributor type object $o \in \mathbf{O^u}$, $\mathbf{u} \preceq \mathbf{a}$, has an extension $\|o\|$ which is a semanteme code: $\|o\| \in \{W^{\star} \mid W \in LEX(\mathbf{u})\}$. E.g., for $o \in L_{\text{BRIGHT}}$, where BRIGHT $\in LEX(\mathbf{a_{grad}})$, $\|o\| = \text{BRIGHT}^{\star}$.

Further, the hierarchy of lexical types induces a hierarchy of the corresponding lexical classes: $\mathbf{u} \preceq \mathbf{v}$ if and only if $L_{\mathbf{u}} \subseteq L_{\mathbf{v}}$ and $L_W \subseteq L_{\mathbf{u}}$ for $W \in LEX(\mathbf{u})$.

Then, all lexical classes L_W representing semantemes $W \in LEX(\mathbf{u})$ share the same attributes. The set of these attributes is denoted $Att(\mathbf{u})$. Every attribute A is characterized by the type $\mathbf{v}$ of its values (notation: $\text{A}^{\mathbf{v}}$). E.g. $\text{DEGREE}^{\mathbf{a_{degr}}}$ with values $(\text{VERY}^{\star})^{\mathbf{a_{degr}}}$, $(\text{SLIGHTLY}^{\star})^{\mathbf{a_{degr}}}$, etc., is an attribute of all classes L_W representing semantemes of type $\mathbf{a_{grad}}$ ($W \in LEX(\mathbf{a_{grad}})$), e.g., $\text{RED}^{\mathbf{a_{grad}}}$, $\text{FAST}^{\mathbf{a_{grad}}}$, etc. It is subsumed that $\mathbf{v} \preceq \mathbf{a}$ ($\mathbf{v}$ is an attributor type) for every attribute $\text{A}^{\mathbf{v}}$. If $Att(\mathbf{u}) = \{\text{A}_1^{\mathbf{v_1}}, \ldots, \text{A}_m^{\mathbf{v_m}}\}$ are all attributes of objects of type $\mathbf{u}$, the set of their value types is denoted $DT(\mathbf{u}) = \{\mathbf{v_1}, \ldots, \mathbf{v_m}\}$. We say that the types in $DT(\mathbf{u})$ are *lexically dependent* on $\mathbf{u}$. It is presumed that the graph of this dependency *has no cycles*. The set of primitive types being finite, this means that there are *minimal* attribute types with no dependents: $DT(\mathbf{a}_0) = \emptyset$. One of minimal types is $\mathbf{a_{degr}}$. Another example is the precision type: $\mathbf{a_{prec}}$ with values $(\text{NEARLY}^{\star})^{\mathbf{a_{prec}}}$, $(\text{ABOUT}^{\star})^{\mathbf{a_{card}}}$, etc., which is the value type of the attribute QUANT of all classes L_W for semantemes of type $\mathbf{a_{card}}$ ($W \in LEX(\mathbf{a_{card}})$), e.g. $\text{TWO}^{\mathbf{a_{card}}}$. We set $o.\text{A} =_{df} \|\text{A}(o)\|$.

In DP semantics, attribute values serve to constrain lexical classes. Here is an example.

Example 3 *The semanteme GLASS in the DP in Fig. 3 has one core* CONTENTS*-argument and two attributes:* $Att(\mathbf{n_{vessel}}) = \{\text{FULLNESS}^{\mathbf{a_{grad}}}, \text{QUANT}^{\mathbf{a_{card}}}\}$. *So the system of constraints for GLASS is defined as* $AC(\text{FULLNESS}^{\mathbf{a_{grad}}} = \pi_1, \text{QUANT}^{\mathbf{a_{card}}} = \pi_2) = AC(\text{FULLNESS}^{\mathbf{a_{grad}}} = \pi_1) \cup AC(\text{QUANT}^{\mathbf{a_{card}}} = \pi_2)$, *where* π_1 *and* π_2 *are the two attributor subplans (*FULLNESS*-branch and* QUANT*-branch) of this DP. Below, in semantics definition, the components are computed bottom-up recursively:* $AC(\text{DEGREE} = \textit{NEARLY}) = \{\text{DEGREE}(o_f) = NEARLY^\star, \|o_f\| = FULL^\star\}$ *for* $o_f = \gamma(\mathsf{z_f})$, $AC(\text{FULLNESS}^{\mathbf{a_{grad}}} = \pi_1) = \{\text{FULLNESS}(o) = o_f\} \cup AC(\text{DEGREE} = \textit{NEARLY})$ *for* $o = \gamma(\mathsf{x})$. *Similar for* $AC(\text{QUANT}^{\mathbf{a_{card}}} = \pi_2)$.

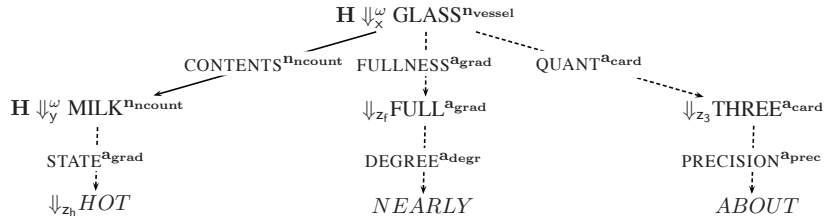


Figure 3. A DP of *about three nearly full glasses of hot milk*.

Finally, the lexical semantics of verbals reduces all verbals' derivatives $V[\mathbf{d}]$ to the canonical form V. For that is used a special product, called *shifted*, allowing to relate their arguments.

Definition 3 *Let* $n > 0$, $d : k \overset{1-1}{\mapsto} n$ *be an argument shift and* $s_1, \ldots, s_n$ *be a sequence of sets. The* shifted product *of this sequence (under shift* d*) is:*

$$\prod_{1 \le i \le k}^{d} s_i =_{df} M_1 \times \ldots \times M_n,$$

where $M_i = s_{d^{-1}(i)}$ *for* $i \in range(d)$ *and* $M_i = \{\perp\}$ *otherwise.*

4 Definition of Static Semantics

In this section we define in parallel the syntax[3] and the static semantics of DP. The correspondence between d- and s-contexts being inessential for this semantics, we fix an s-context $\sigma =< \Gamma, \Lambda, \Theta, H >$ in which, for every DP π, will be defined its *s-extension* $\|\pi\|^\sigma$. So Γ is a global assignment, Λ is a local assignment, Θ is a nominal objects' evaluation and H is a horizon line. As we shall see, every composite subplan π of a DP is uniquely identified by a global reference x introduced by a determiner: $\pi = \mathbf{D}_{\mathsf{x}}\pi'$. The static semantics $\|\pi\|^\sigma$ will be defined through the extension $\|\Gamma(\mathsf{x})\|^\sigma$ of the object $\Gamma(\mathsf{x})$.

I. Primitives.

I.1. Lexical classes. For a non-attributor semanteme W, $\|L_W\|^\sigma = H(L_W)$.

I.2. Null plans (intuitively, corresponding to existentially bound arguments).

For a null nominal plan $\pi = \Downarrow_{\mathsf{x}} 0^{\mathbf{n}}$ (**Ex:** *Testamentary succession*$_{\text{OBJ}:\Downarrow_{\mathsf{x}} 0^{\mathbf{n}}}$ *goes to Mary*),
$\|\pi\|^\sigma = \|\Gamma(\mathsf{x})\|^\sigma$, where $\|\Gamma(\mathsf{x})\|^\sigma = \{\perp\}$.

For a null attributor plan $\pi = \Downarrow_{\mathsf{x}} 0^{\mathbf{a}}$ (**Ex:** *happy*$_{\text{DEGREE}:0^{\mathbf{a_{degr}}}}$ *as goblin*),
$\|\pi\|^\sigma = \|\Gamma(\mathsf{x})\|^\sigma$, where $\|\Gamma(\mathsf{x})\|^\sigma = \perp$.

I.3. Shifter plans. Let $\pi = \Downarrow_{\mathsf{x}} (\text{K}^{\mathbf{n}'})_{\mathbf{sh}}$, where $(\text{K}^{\mathbf{n}'})_{\mathbf{sh}}$ is a nominal shifter constant of type $\mathbf{n}'$ (**Ex:** $(\mathbf{speaker}^{\mathbf{n_a}})_{\mathbf{sh}}$, $(\text{John}^{\mathbf{n_a}})_{\mathbf{sh}}$). Then:
$\|\pi\|^\sigma = \|\Gamma(\mathsf{x})\|^\sigma$, where $\|\Gamma(\mathsf{x})\|^\sigma = \{((\text{K})_{\mathbf{sh}})^\star\}$.

I.4. Reference plans.

$\|\pi\|^\sigma = \|\Lambda(u)\|^\sigma$ for $\pi = u^{\mathbf{t}}$, u being a local reference.
$\|\pi\|^\sigma = \|\Gamma(\mathsf{x})\|^\sigma$ for $\pi = \Uparrow_{\mathsf{x}^{\mathbf{t}}}$, x being a global reference.

I.5. Primitive attributor plans. $\pi = W$, where $W \in LEX(\mathbf{v})$ and $\mathbf{v} \preceq \mathbf{a}$ is a minimal attributor type (e.g. $\mathbf{a_{degr}}$, $\mathbf{a_{prec}}$), are the only nonreferenced DP. For such DP, $\|\pi\|^\sigma = W^\star$.

II. Compound DP.

Sentential plans.

[3] Because of space limits we omit the rules of visibility of references.

II.1. Unit sentential plans. Let $\pi = \Downarrow_{\mathsf{x}} V[\mathbf{d}](\mathrm{R}_1 : \pi_1, \ldots, \mathrm{R}_k : \pi_k, \mathrm{A}_1 : \pi'_1, \ldots, \mathrm{A}_m : \pi'_m)$ be a sentential DP in which $\pi'_i = \Downarrow_{\mathsf{x}_i} \pi''_i$, $1 \leq i \leq m$, are composite attributor DP. Then:
$\|\pi\|^\sigma = \|\Gamma(\mathsf{x})\|^\sigma$,
$\|\Gamma(\mathsf{x})\|^\sigma = \prod\limits_{1 \leq i \leq k}^{\mathbf{d}} \|\pi_i\|^\sigma$.
$\Gamma(\mathsf{x}) \in \|(L_{\mathbf{V}})\|^\sigma$, $\mathrm{A}_i(\Gamma(\mathsf{x})) = \Gamma(\mathsf{x}_i)$ and $\Gamma(\mathsf{x}).\mathrm{A}_i = \|\pi'_i\|^\sigma$, $1 \leq i \leq m$.

II.2. Coordinated sentential plans. Let $\pi = \Downarrow_{\mathsf{x}} \mathcal{C}^{(n)}(\pi_1, \ldots, \pi_n)$, where $n > 1$ and $\pi_i = \Downarrow_{\mathsf{x}_i} \pi'_i$ are unit sentential DP of sentential types $\mathbf{s_i}$, $1 \leq i \leq n$. Then:
$\|\pi\|^\sigma = \|\Gamma(\mathsf{x})\|^\sigma$, where $\|\Gamma(\mathsf{x})\|^\sigma = <\Gamma(\mathsf{x}_1), \ldots, \Gamma(\mathsf{x}_n)>$,
$\|\Gamma(\mathsf{x}_i)\|^\sigma = \|\pi'_i\|^\sigma, 1 \leq i \leq n$.

Nominal plans.

II.3. Absolute unit determined nominal plans. Let $\pi = \mathbf{D}_{\mathsf{x}}\hat{\pi}$, where $\mathbf{D}_{\mathsf{x}} = (\mathbf{Q} \Downarrow^{\mathbf{k}}_{\mathsf{x}} u)$ is a determiner in which $\mathbf{Q} \in \{\mathbf{H}, \mathbf{I}\}$, $\mathsf{x}^{\mathbf{n}'}$ is a global reference, u is a local reference, $\mathbf{k}$ is a number or ω, $N^{\mathbf{t}}$ is a nominal of type $\mathbf{t} = (\mathrm{S}_1^{\mathbf{n_1}} \ldots \mathrm{S}_k^{\mathbf{n_k}} \mathrm{A}_1^{\mathbf{v_1}} \ldots \mathrm{A}_m^{\mathbf{v_m}} \rightarrow \mathbf{n}')$, $\hat{\pi} = N^{\mathbf{t}}(\mathrm{S}_1 : \pi_1, \ldots, \mathrm{S}_k : \pi_k, \mathrm{A}_1 : \pi'_1, \ldots, \mathrm{A}_m : \pi'_m)$ is a determinerless nominal DP, where $\mathbf{n}' \preceq \mathbf{n}$, $\pi_i = \mathbf{D}_{\mathsf{x}_i}\hat{\pi}_i$, $1 \leq i \leq k$, are core argument nominal DP and $\pi'_j = \Downarrow_{\mathsf{y}_j} \hat{\pi}'_j$, $1 \leq j \leq m$, are composite attributor DP (see the DP in Fig. 3 and Example 3). Then:
$\|\pi\|^\sigma = \{\Gamma(\mathsf{x})\}$, if $\mathbf{Q} = \mathbf{H}$, and $\|\pi\|^\sigma = \|\Gamma(\mathsf{x})\|^\sigma$, if $\mathbf{Q} = \mathbf{I}$,
$\|\Gamma(\mathsf{x})\|^\sigma = \Theta(\Gamma(\mathsf{x}))$ and $card(\|\Gamma(\mathsf{x})\|^\sigma) \leq \mathbf{k}$,
$\Gamma(\mathsf{x}) \in \|(L_N)\|^\sigma$,
$\mathrm{S}_i(\Gamma(\mathsf{x})) = \Gamma(\mathsf{x}_i)$, $1 \leq i \leq k$.
$\mathrm{A}_j(\Gamma(\mathsf{x})) = \Gamma(\mathsf{y}_j)$ and $\Gamma(\mathsf{x}).\mathrm{A}_j = \|\pi'_j\|^\sigma$, $1 \leq j \leq m$.

II.4. Relativized unit determined nominal plans. Let $\pi = \mathbf{D}_{\mathsf{x}}\pi_1$, where $\mathbf{D}_{\mathsf{x}} = (\mathbf{Q} \Downarrow^{\mathbf{k}}_{\mathbf{xry}} u)$ is a determiner in which $\mathbf{Q} \in \{\mathbf{H}, \mathbf{I}\}$, $\mathbf{r} \in \{\dot{\in}, \sim, \subset, /, \ldots\}$, π_1 is a determinerless nominal plan and y is a global object reference identifying in the preceding discourse a nominal plan $\mathbf{D}_{\mathsf{y}}\pi_0$ with determiner $\mathbf{D}_{\mathsf{y}} = (\mathbf{Q_0} \Downarrow^{\mathbf{k_0}}_{\mathsf{y}} u_0)$ (see the DP in Fig. 2). Then $\|\pi\|^\sigma$ is defined as in the preceding case. Besides this, the following $\mathbf{r}$-conditions also hold:
$\|\mathbf{r}\|^\sigma(\Gamma(\mathsf{x}), \Gamma(\mathsf{y}))$ if $\mathbf{r} \in \{\sim, \subset, /, \ldots\}$,
$\Gamma(\mathsf{x}) \in \|\Gamma(\mathsf{y})\|^\sigma$, $\Lambda(u) = \{\Gamma(\mathsf{x})\}$ and $card(\|\Gamma(\mathsf{y})\|^\sigma) \leq \mathbf{k_0}$ if $\mathbf{r} = \dot{\in}$.

II.5. Relative determined nominal plans. Let $\pi = \iota_{\mathrm{R}}(\pi_0 \mid \hat{\pi}_0)$, where $\pi_0 = \mathbf{D}_{\mathsf{x}}\pi'_0$ is a unit determined nominal plan, u is the local reference in $\mathbf{D}_{\mathsf{x}}$, R is a role and $\hat{\pi}_0 = \Downarrow_{\mathsf{y}} V[\mathbf{d}](\mathrm{R}_1 : \hat{\pi}_1, \ldots, \mathrm{R}_i : u, \ldots, \mathrm{R}_k : \hat{\pi}_k, \mathrm{A}_1 : \hat{\pi}'_1, \ldots, \mathrm{A}_m : \hat{\pi}'_m)$ is a sentential plan such that $\mathrm{R}_i = \mathrm{R}$. Let also $I^\sigma_{\mathrm{R}}(\hat{\pi}_0) = \{x =< x_1, \ldots, x_k > \mid (\exists y_1, \ldots, y_n)(< y_1, \ldots, y_n > \in \|\Gamma(\mathsf{y})\|^\sigma \ \& \ x_i = y_{d^{-1}(i)})\}$. Then:
$\|\Gamma(\mathsf{x})\|^\sigma = \|\pi_0\|^\sigma \cap I^\sigma_{\mathrm{R}}(\hat{\pi}_0)$ and
$\|\pi\|^\sigma = \{\Gamma(\mathsf{x})\}$, if $\mathbf{Q} = \mathbf{H}$, and $\|\pi\|^\sigma = \|\Gamma(\mathsf{x})\|^\sigma$, if $\mathbf{Q} = \mathbf{I}$.
If π_0 is a relativized unit determined nominal plan (i.e. $\mathbf{D}_{\mathsf{x}} = \mathbf{Q} \Downarrow^{\mathbf{k}}_{\mathbf{xry}} u$), then the $\mathbf{r}$-conditions also hold.
Ex: Relative and comparative clauses.

II.6. Aggregate nominal plans. Let $\pi = \mathbf{D}_{\mathsf{x}}\mathcal{A}(\pi_1, \ldots, \pi_n)$, where $\mathbf{D}_{\mathsf{x}} = \mathbf{Q} \Downarrow^{\mathbf{k}}_{\mathsf{x}} u$ and $\pi_i = \mathbf{D}_{\mathsf{x}_i}\pi'_i$, $1 \leq i \leq n$, are determined nominal DP. Then:
$\|\Gamma(\mathsf{x})\|^\sigma = \Theta(\Gamma(\mathsf{x}))$ and $\Gamma(\mathsf{x}_1), \ldots, \Gamma(\mathsf{x}_n) \in \|\Gamma(\mathsf{x})\|^\sigma$ if $\mathbf{k} = \omega$,
$\|\Gamma(\mathsf{x})\|^\sigma = \{\Gamma(\mathsf{x}_1), \ldots, \Gamma(\mathsf{x}_n)\}$ if $\mathbf{k} = n$,
$\|\pi\|^\sigma = \{\Gamma(\mathsf{x})\}$, if $\mathbf{Q} = \mathbf{H}$, and $\|\pi\|^\sigma = \|\Gamma(\mathsf{x})\|^\sigma$, if $\mathbf{Q} = \mathbf{I}$.
Ex: *(Students$_{\Downarrow_{\mathsf{x}_1}}$ and professors$_{\Downarrow_{\mathsf{x}_2}}$)$_{\Downarrow_{\mathsf{x}}}$ went on strike.*

Attributor plans.

II.7. Lexicalized attributor plans. Let $\pi = \Downarrow_{\mathsf{x}} W^{\mathbf{t}}(\mathrm{A}_1 : \pi_1, \ldots, \mathrm{A}_m : \pi_m)$ be a DP in which $\mathbf{t} = (\mathrm{A}_1^{\mathbf{v_1}} \ldots \mathrm{A}_m^{\mathbf{v_m}} \rightarrow \mathbf{u})$, $\mathbf{u} \preceq \mathbf{a}$, and π_i are attributor DP, $1 \leq i \leq m$. Then:
$\|\pi\|^\sigma = \|\Gamma(\mathsf{x})\|^\sigma = W^\star$,

$\text{A}_i(\Gamma(\mathsf{x})) = \Gamma(\mathsf{x}_i)$, if $\pi_i = \Downarrow_{\mathsf{x}_i} \pi'_i$, and $\text{A}_i(\Gamma(\mathsf{x})) = \|\pi_i\|^\sigma$ otherwise, $1 \leq i \leq m$,
$\Gamma(\mathsf{x}).\text{A}_i = \|\pi_i\|^\sigma$, $1 \leq i \leq m$.
Ex: See the DP in Fig. 3 and Example 3.
II.8. Relative attributor plans. Let $\pi = \iota\left(\Downarrow_{\mathsf{x}} 0^{\mathbf{t}} \mid \pi_1\right)$ be a relative attributor plan in which $\mathbf{u} \preceq \mathbf{a}$ is an attributor type, $\mathsf{x}^{\mathbf{u}}$ is a global reference and $\pi_1 = \Downarrow_{\mathsf{y}} \pi'_1$ is a sentential type DP. Then:
$\|\pi\|^\sigma = \|\Gamma(\mathsf{x})\|^\sigma = \bot$,
$\|\pi_1\|^\sigma = \|\Gamma(\mathsf{y})\|^\sigma$,
$\|\mathbf{rel}\|^\sigma(\Gamma(\mathsf{x}), \Gamma(\mathsf{y}))$ for a special relation $\mathbf{rel}$.
Ex: *He was so*$_{\Downarrow_{\mathsf{x}}(0^{\mathbf{adegr}})}$ *glad, that ...*

5 On Dynamic DP Semantics

The dynamic semantics is defined through translation$\lceil .. \rceil$: for a discourse $\delta = (\pi_1, \ldots, \pi_n)$, $\lceil \delta \rceil = \lceil \pi_1 \rceil \ldots \lceil \pi_n \rceil$ is a reaction-to-stimuli process which, when applied to a starting context Σ_0, executes transitions $(\pi_1)^{\Sigma_1}_{\Sigma_0}, (\pi_2)^{\Sigma_2}_{\Sigma_1}, \ldots, (\pi_n)^{\Sigma_n}_{\Sigma_{n-1}}$ and computes the corresponding d-extensions: $|\delta|_{\Sigma_0} = (|\pi_1|^{\Sigma_1}_{\Sigma_0}, |\pi_2|^{\Sigma_2}_{\Sigma_1}, \ldots, |\pi_n|^{\Sigma_n}_{\Sigma_{n-1}})$. The translation and the transition actions and rather technical and will be published elsewhere. Here we only illustrate it by the process corresponding to the discourse in Fig. 1,2.

Its intermediate data are shown in Tables 1,2 with columns: Context (current d-context), GRef (global object reference identifying a subplan), Oid (identity of the created object), d-Extension elements (elements added to the d-extension of the object), LRef (local object reference), LVal (current value of the local object reference), Attributes (attribute value extension) and Semanteme (the root semanteme of the subplan). This computation executes two processes: $\lceil \pi_1 \rceil = p_1$ (see Table 1) and $\lceil \pi_2 \rceil = p_2$ (see Table 2), where π_1 and π_2 are DP in Figures 1 and 2 respectively.

Context	GRef	Oid	d-Extension elements	LRef	LVal	Attributes	Semanteme
Σ_0	x_v	$o_v \in \mathbf{O^n}$	$(\text{v})^\star_{\mathbf{sh}}$				VILLAGE
Σ_1	x_f	$o_f \in \mathbf{O^{n_a}}$	$\bot$	u_f	$\{\bot\}$		FARMER
Σ_2	s_1	$o_{loc} \in \mathbf{O^s}$	$<$ DFM : $\bot$, DFS : $o_v >$				**loc**
Σ_3	x_t	$o_t \in \mathbf{O^n}$	$\bot$	u_t	$\{\bot\}$		TRACTOR
Σ_4	s_2	$o_u \in \mathbf{O^s}$	$<$ SIA : $\bot$, OIP : $\bot >$			o_u.PSTATUS = DCL*, etc.	USE
Σ_5	x_n	$o_n \in \mathbf{O^{n_a}}$	$\bot$	u_n	$\{\bot\}$		NEIGHBOR
Σ_6	s_3	$o_{sh} \in \mathbf{O^s}$	$<$ SIA : $\bot$, OIP : $\bot$, CO-SIA : $\bot >$			o_u.PSTATUS = DCL*, etc.	SHARE
Σ_7	s_4	$o_h \in \mathbf{O^s}$	$<$ SIA : $\bot$, OIP : $\bot >$			o_h.PSTATUS = DCL*, etc.	HAVE

Table 1. Computation for the first DP π_1.

Context	GRef	Oid	d-Extension elements	LRef	LVal	Attributes	Semanteme
Σ_8	y_f	$o_B \in \mathbf{O^{n_a}}$	$(\text{Bob})^\star_{\mathbf{sh}}$	u'_f	$\{o_B\}$		$(\text{Bob})_{\mathbf{sh}}$
	x_f	o_f	$\bot, o_B$	u_f	$\{o_B\}$		
	s_2	o_u	$<$ SIA : o_B, OIP : $\bot >$				USE
	s_4	o_h	$<$ SIA : o_B, OIP : $o_n >$				HAVE
Σ_9	y_t	$o_{hv} \in \mathbf{O^n}$	$\bot$	u'_t	$\{o_{hv}\}$	o_{hv}.STATE = OLD*	HARVESTER
	x_t	o_t	$\bot, o_{hv}$	u_t	$\{o_{hv}\}$		TRACTOR
	s_1	o_u	$<$ SIA : o_B, OIP : $o_{hv} >$				USE
Σ_{10}	s_5	$o_{appurt} \in \mathbf{O^s}$	$<$ DFM : o_{hv}, DFS : $o_B >$				**appurt**
Σ_{11}	y_n	$o_T \in \mathbf{O^{n_a}}$	$(\text{Tom})^\star_{\mathbf{sh}}$	u'_n	$\{o_T\}$		$(\text{Tom})_{\mathbf{sh}}$
	x_n	o_n	$\bot, o_T$	u_n	$\{o_T\}$		NEIGHBOR
	s_3	o_{sh}	$<$ SIA : o_B, OIP : o_{hv}, CO-SIA : $o_T >$				SHARE
Σ_{12}	s_6	o'_{sh}	$<$ SIA : o_B, OIP : o_{hv}, CO-SIA : $o_T >$			o'_{sh}.PSTATUS = DCL*, etc.	SHARE

Table 2. Computation corresponding to the second DP π_2.

This computation executes two processes: $\lceil \pi_1 \rceil = p_1$ and $\lceil \pi_2 \rceil = p_2$, where π_1 and π_2 are DP in Figures 1 and 2 respectively. The computation of p_1 (see Table 1) is started in context Σ_0 in which there is an object $o_v = \gamma_{\Sigma_0}(\mathsf{x}_v)$ referenced by $\Uparrow_{\mathsf{x}_v}$ (*the village*). Then, in the course of seven consecutive transitions, it creates a new object o for each subplan identified by its determiner $\mathbf{D}_{\mathsf{x}}$, binds the global reference x with o and adds the object to the accessible subset $\hbar_{\Sigma_1}(L_W)$ of the lexical class L_W corresponding to the head semanteme W of the subplan. In the case where W is a verbal, the process adds new facts to the shifted product related with L_W (the element of the column "Semanteme" identifies W and the corresponding subplan). For instance, the transition to Σ_1 is due to the determiner $(\mathbf{I} \Downarrow_{\mathsf{x}_f} u_f)$ applied to FARMER. The process creates a new object $o_f = \gamma_{\Sigma_1}(\mathsf{x}_f) \in \mathbf{O}^{\mathbf{n_a}}$ with *uncertain* extension $\{\bot\}$ which becomes in context Σ_1 the d-extension of the subplan with the head semanteme FARMER. The object o_f is added to $\hbar_{\Sigma_1}(L_{FARMER})$ and $\lambda_{\Sigma_1}(u_f)$ is set to $\{\bot\}$. In the next transition to Σ_2, due to the determiner $\Downarrow_{\mathsf{s}_1}$, the process creates a new object o_{loc} for the proper relation $\mathbf{loc}$, binds s_1 with o_{loc} and adds the fact $< \text{DFM} : \bot, \text{DFS} : o_v >$ to the shifted product representing its extension. Importantly, in the next transition to Σ_3, the process, unlike the transition to Σ_1, creates an object $o_t \in \mathbf{O}^{\mathbf{n}}$ for TRACTOR with the *certain* extension $\{o_t\}$. This is explained by the difference of access constructors in the two determiners: *individual* $\mathbf{I}$ for FARMER and *holistic* $\mathbf{H}$ for TRACTOR. This difference manifests itself in the computation of p_2 (see Table 2). Viz., due to the determiner $(\mathbf{H} \Downarrow^1_{\mathsf{y}_f \in \mathsf{x}_f} u'_f)$ applied to $(\text{Bob}^{\mathbf{n_a}})_{\mathbf{sh}}$, this computation changes Σ_7 to Σ_8, creates $o_B = \gamma_{\Sigma_8}(\mathsf{y}_f)$ $\in \mathbf{O}^{\mathbf{n_a}}$ with extension $\theta_{\Sigma_8}(o_B) = \{(\text{Bob})^{\star}_{\mathbf{sh}}\}$, and, due to the relativized reference $\mathsf{y}_f \dot{\in} \mathsf{x}_f$, raises a stimulus to which reacts the object o_f binding the reference x_f. The reaction consists in reactivation of the process p_1 which adds o_B to the extension $|o_f|^{\Sigma_8}$ and to $\hbar_{\Sigma_8}(L_{FARMER})$, and binds the local reference u_f with $\{o_B\}$ ($\lambda_{\Sigma_8}(u_f) = \{o_B\}$), whereby the shifted products for USE and HAVE are recomputed: $< \text{SI}\mathbf{A} : o_B, \text{OI}\mathbf{P} : \bot >$ is added to the former and $< \text{SI}\mathbf{A} : o_B, \text{OI}\mathbf{P} : o_n >$ is added to the latter. A similar effect is seen later in the transitions to Σ_9 and to Σ_{11}.

This illustration explains the difference, in DP semantics, between the nominal DP $\pi_1 = (\mathbf{H} \Downarrow^{\mathbf{k}_1}_{\mathsf{x}\ \phi} u_1)N_1(\overline{\mathsf{A} : \pi})$ with holistic determiner and the nominal DP $\pi_2 = (\mathbf{I} \Downarrow^{\mathbf{k}_2}_{\mathsf{y}\ \psi} u_2)N_2(\overline{\mathsf{A} : \pi})$ with individual determiner. The former has *invariant* certain extension $|o_1|^{\Sigma} = \{o_1\}$ in which o_1 is the object which realizes π_1 in the discourse and binds the reference $\mathsf{x} : \gamma_{\Sigma}(\mathsf{x}) = o_1$ (cf. point **II.3.** of the definition of static DP semantics in the case of $\mathbf{Q} = \mathbf{H}$). The latter has a set extension $|o_2|^{\Sigma}$, where $\gamma_{\Sigma}(\mathsf{y}) = o_2$, evolving in the discourse. Viz., every time the DP π_2 is referred in the discourse by another DP, say π_1, through the relativized reference $\mathsf{x}\ \phi = \mathsf{x} \dot{\in} \mathsf{y}$, its extension $|o_2|^{\Sigma}$ is *updated*: $|o_2|^{\Sigma_1} = |o_2|^{\Sigma} \cup \{o_1\}$, as well as its local variable: $\lambda_{\Sigma_1}(u_2) = \{o_1\}$, the predications of the verbals for which π_2 is an argument, either directly (cf. point **II.3.** of the definition of static DP semantics in the case of $\mathbf{Q} = \mathbf{I}$), or relatively, through u_2 (cf. point **II.5.** of the definition of static DP semantics), are *recomputed*: the facts with the new witness o_1 are added to their shifted product. The determiner's parameter $\mathbf{k}$ stands for the intended cardinality of the nominal object extension. In particular, $\mathbf{k} = 1$ corresponds to the *singular* and ω imposes no constraints on the cardinality. In this way is expressed in dynamic DP semantics its specific *plurality-through-evidence*: only the entities mentioned in the discourse are added to nominal extensions and only the facts witnessed by such entities emerging in the discourse are added to verbal extensions.

In principle, DP-determiners may use rather complex relations constraining objects in the extension of nominals. For instance, the determiner $\mathbf{D}_c = \big(\mathbf{H} \Downarrow^{\omega}_{\mathsf{x}} \big(\mathbf{H} \Downarrow^{\omega}_{\mathsf{y}} \text{ROSE}, \mathbf{H} \Downarrow^{\omega}_{\mathsf{z}} \text{LILIES}\big)\ (\mathbf{card}(\mathsf{y}) > \mathbf{card}(\mathsf{z}))\big)$ will be used in the OI**P**-subplan $\mathbf{D}_c\mathcal{A}_{\cup}\{\Uparrow_{\mathsf{y}}, \Uparrow_{\mathsf{z}}\}$ of *At least three girls gave* $(\textit{more roses than lilies})_{\mathbf{D}_c}$ *to John*. Such determiners make them, in practice, comparable with so called cumulative quantifiers generally treated using generalized quantifiers (cf. (Keenan, 1996; Keenan and Westerståhl, 1997)). In our example, $\mathcal{A}_{\cup}\{\Uparrow_{\mathsf{y}}, \Uparrow_{\mathsf{z}}\}$ is a nominal aggregate with union extension: $\big|\mathcal{A}_{\cup}\{o_1, o_2\}\big|^{\Sigma} = |o_1|^{\Sigma} \cup |o_2|^{\Sigma}$. By the way, among the constraints imposed by the determiners, there is the *co-reference constraint* $\mathsf{x} \sim \mathsf{y}$ saying that the (different) objects $\gamma(\mathsf{x})$ and $\gamma(\mathsf{y})$ *represent the same entity*, as it is the case in the discourse:

$(Lincoln)_{(\mathbf{H}\Downarrow^1_{\mathsf{x}}u)(\text{LINCOLN})_{\mathbf{sh}}}$ *was born in 1809.* $(This\ President)_{(\mathbf{H}\Downarrow^1_{\mathsf{y}\sim\mathsf{x}}v)PRESIDENT}$ *was a liberal*. These constraints directly correspond to the co-reference in the logical dynamic semantics such as DRT (Kamp et al.,). The difference is that in DP semantics the co-reference is not checked.

On the other hand, an object o_1 satisfying the determiners' constraints gets to the set-extension of a nominal object o_2 only through the reaction to the effective stimulus corresponding to a DP π_1 referring π_2 in the discourse. So from the point of view of extension constraints, the individual determiners are rather close to the universal quantifier in the first order logics. At the same time, they are very different from $\forall$ because of this plurality-through-evidence interpretation. As to the holistic determiners, they were always a problem to express in the traditional logical semantics. They allow to adequately express the meaning of noun phrases as in *John likes* $(books)_{(\mathbf{H}\Downarrow^\omega_{\mathsf{x}}u)BOOK}$ and provide a holistic interpretation for mass nominals as in *He needs more* $(water)_{\mathbf{H}\Downarrow^0_{\mathsf{x}}WATER}$.

Main property of DP semantics. We show that the dynamic and the static DP semantics coincide in the corresponding dynamic and static contexts.

Theorem 1 *Let* $\delta = (\pi_1, \ldots, \pi_n)$ *be a discourse,* Σ_0 *be an initial d-context,* $|\delta|_{\Sigma_0} = (|\pi_1|^{\Sigma_1}_{\Sigma_0}, |\pi_2|^{\Sigma_2}_{\Sigma_1}, \ldots, |\pi_n|^{\Sigma_n}_{\Sigma_{n-1}})$ *be the d-semantics of* δ *relative to* Σ_0 *and* $\sigma_i = < \gamma_{\Sigma_i}, \lambda_{\Sigma_i}, \theta_{\Sigma_i}, \hbar_{\Sigma_i} >$ *be the s-contexts corresponding to d-contexts* Σ_i*. Then* $|\pi_i|^{\Sigma_i}_{\Sigma_{i-1}} = \|\pi_i\|^{\sigma_i}$ *for all* $i, 0 < i \leq n$.

6 Conclusion

One can see that the speaker's stance discourse semantics outlined in this paper has not much to do with the Grice's implicatures (Grice, 1989). Nor has it something to do with processing of hearer's beliefs depending on an interpretation of speaker's discourse. It is also very different from all logical DRT-like semantics of discourse (cf. (Heim, 1983; Kamp and Reyle, 1993; Kamp et al., ; Muskens et al., 1997)). The anaphora resolution, the emblem of logical discourse representation theories, is not included into the DP semantics because the referential relations are explicitly marked in DP using its determiners. Some of these referential relations established by these determiners, such as co-reference $\sim$, and attribute value comparison relations $<, >, =$, as well as the signs $+, -$ of verbal objects may introduce conflicts in the contexts. Checking of probable inconsistencies in the contexts is not required in DP semantics. This makes possible to apply it to correctly constructed DP with contradictory meaning, which is impossible in all kinds of logical theories of discourse. By the way, these special features make the DP semantics efficiently implementable. It has a polynomial time complexity (the consistency check included).

Due to object-orientation, the DP semantics goes without quantifiers. At that, there are certain similarities between the conventional quantifiers and the DP determiners. Creation of an object ($\gamma(\mathsf{x}) = o$) is an analog of the existential quantifier. It is closer than the logical quantifier to the natural language "existence": *every entity mentioned in the discourse exists*. The object access connectors $\mathbf{H}$ and $\mathbf{I}$ correspond to two different concepts of universal quantification. The former, holistic, has no analogues in the traditional logics, the latter, individual, is rather close to $\forall$ from the point of view of extension constraining: the cumulative determiners $((\mathbf{H} \Downarrow_{\mathsf{x}_1})N_1, \ldots, (\mathbf{H} \Downarrow_{\mathsf{x}_n})N_n)_{\mathbf{r}(\mathsf{x}_1,\ldots,\mathsf{x}_n)}$ with unlimited relations $\mathbf{r}$ are not less expressive than the cumulative quantifiers used in the plurality constraints definitions in terms of generalized quantifiers. At the same time, the individual determiners express a specific plurality-through-evidence. In the end, it is due to this "quantifier-freeness" that the static DP semantics is fully compositional (DP-determiners are interpreted *in situ*, i.e. in verbals' argument positions).

The DP semantics is a kind of formal semantics fitting well the Meaning-Text Theory frame, because it applies to a meaning structure designed for discourse generation and does not require consistency of the meaning structures to which it applies. Verbals' diathetic shifts make DP very flexible and well adapted to the traditional linguistic semantical representations. In fact, they are very close to those introduced and studied by E. Paducheva (see (Padučeva, 2003; Padučeva, 2004)). The use of communicative ranks in

definitions of argument shifts allows to express some aspects of communicative structure. To our knowledge, it is the first formal semantics taking in consideration diathetic shifts in predication.

Due to the interpretation of non-core arguments as representing constraints on the attribute values, the semantical function-argument dependencies in DP semantics do not conflict with the natural surface syntactic dependencies. For instance, in DP, attributor type semantemes are arguments of nominals, which reflects the surface dependency of modifiers on the modified nouns (to compare with the conventional logical semantics, in which, quite the contrary, a nominal object is the argument of the property expressing a noun's modifier). This structural conformity has an exact form: in (Dikovsky, 2007) we show how syntactic categorial dependency grammar types may be generated from DP by finite tree transducers. This transduction may be seen as a formal model for the Meaning-Text Theory.

References

Dikovsky, Alexander and Boris Smilga. 2005. Semantic roles and diatheses for functional discourse plans. In *Proc. of the 2nd International Conference "Meaning-Text Theory" (MTT 2005)*, pages 98–109.

Dikovsky, Alexander. 2003. Linguistic meaning from the language acquisition perspective. In *Proc. of the 8th Intern. Conf. "Formal Grammar 2003" (FG 2003)*, pages 63–76, Vienna, Austria, August.

Dikovsky, A. 2007. A finite-state functional grammar architecture. In Leivant, D. and Ruy de Queiroz, editors, *Logic, Language, Information and Computation. Proc. of the 14th Intern. Workshop WoLLIC 2007*, LNCS 4576, pages 131–146, Rio de Janeiro, Bresil. Springer Verlag.

Grice, H.P. 1989. *Studies in the Way of the Words.* Harvard University Press, Cambridge, MA.

Heim, I. 1983. File change semantics and the familiarity theory of definiteness. In Bäuerle, R., C. Schwarze, and von A. Stechow, editors, *Meaning. Use and Interpretation of Language*. De Gruyter, Berlin.

Kamp, Hans and Uwe Reyle. 1993. *From Discourse to Logic: An introduction to modeltheoretic semantics, formal logic and Discourse Representation Theory.* Kluwer Academic Publishers, Dordrecht, Germany.

Kamp, Hans, Josef van Genabith, and Uwe Reyle. Discourse representation theory. In Gabbay, D.M. and F. Guenthner, editors, *Handbook of Philosophical Logic*. Forthcoming, Second edition.

Keenan, E. and D. Westerståhl. 1997. Generalized quantifiers in linguistics and logic. In van Benthem and ter Meulen (van Benthem and ter Meulen, 1997), chapter 15, pages 837–893.

Keenan, E. 1996. The semantics of determiners. In Lappin, S., editor, *The Handbook of Contemporary Semantic Theory*, pages 41–63. Blackwell.

Muskens, R., J. van Benthem, and A. Visser. 1997. Dynamics. In van Benthem and ter Meulen (van Benthem and ter Meulen, 1997), chapter 10, pages 587–648.

Padučeva, Elena V. 2003. Diathesis: some extended applications of the term. In *Proc. of the 1st Intern. Conf. on Meaning-Text Theory*, pages 201–210, Paris, ENS, June 16–18.

Padučeva, Elena V. 2004. *Dynamic models in lexical semantics.* Jazyki slavjanskoj kul'tury, Moscow. (Russ.).

van Benthem, J. and A. ter Meulen, editors. 1997. *Handbook of Logic and Language*. North-Holland Elsevier, The MIT Press, Amsterdam, Cambridge.

YES and NO:
Universal Ideas in Language Specific Configurations[1]

Dmitrij Dobrovol'skij
Russian Language Institute,
Russian Academy of Sciences
Volkhonka 18/2
119019 Moscow
dm-dbrv@yandex.ru;
Dmitrij.Dobrovolskij@assoc.oeaw.ac.at

Irina Levontina
Russian Language Institute,
Russian Academy of Sciences
Volkhonka 18/2
119019 Moscow
irina.levontina@mail.ru

Abstract

Even the most general and seemingly universal ideas have a unique representation in every single language. In this paper, we analyze the Russian words *да* 'yes' and *нет* 'no', as well as their English and German correlates. We show that these words behave differently from language to language. As empirical basis, we took two Harry Potter books by J. K. Rowling (*Harry Potter and the Sorcerer's Stone* and *Harry Potter and the Chamber of Secrets*) and their German and Russian translations. Many of the cross-linguistic differences discussed in the paper go back to the language specific properties of the discourse structure.

1 Introduction

It is well-known that even the most general and seemingly universal ideas have a unique representation in every single language. In this paper, we analyze the Russian words *да* 'yes' and *нет* 'no', as well as their English and German correlates. We will show that these words behave differently from language to language even in their central readings. Let us start with an example.

In 2008 Barack Obama won the United States presidential election under the slogan "Yes, we can!". From the linguistic point of view, it is remarkable that this seemingly so simple a sentence cannot be adequately translated into Russian, at least not in all possible contexts. The stumbling block here is the word *yes*. In English, we can easily imagine a dialogue, such as – *You cannot do it! – Yes, I can*. Its literal translation into Russian would be quite unacceptable. Cf. – *Вы этого не можете! – Да, могу*. The correct answer here is: – *Нет, могу*. Incidentally, in German it is also not possible to use the word *ja* 'yes' in such cases. There is a special word *doch* (something like the combination *нет да* in Russian). Compare *Doch, ich kann es.*

Hence, Obama's formula *Yes, we can!* cannot be translated into Russian by *Да, мы можем*. In certain cases, the correct translation would be *Нет, мы можем*. Compare passages like *They say, we are not ready, we cannot. Yes, we can!* The literal translation *Нам говорили, что мы не готовы, что мы не можем. Да, мы можем* would lack any coherence. However, if a refrain of this kind was translated differently from context to context it would lose its slogan nature. The simplest way to avoid such difficulties is to omit *yes* translating *Yes, we can!* by *Мы можем!* But the *yes* at the beginning of the sentence is very important. First, it makes the utterance dialogic, and second, it carries a strong positive attitude. Hence, to simply omit *yes* is not possible.

What is important is not only the question whether or not the given words are equivalent in principle, but as to how they really function in discourse. This question cannot be answered without using parallel corpora. It is obvious that using as empirical data contexts from various Russian, English and German

[1] Acknowledgements and thanks are due to anonymous referees for comments on early drafts. Work at various stages was supported by RGNF-grant 08-04-00173a.

text sources which are not identical in content would obscure the results because many other factors play a role in this case and it is not possible to take all of them into account. Also self-constructed sentences as linguistic data have obvious restrictions. Thus parallel corpora seem to be the best starting point for our cross-linguistic analysis. As an empirical basis, we took two Harry Potter books by J. K. Rowling (*Harry Potter and the Sorcerer's Stone* and *Harry Potter and the Chamber of Secrets*) with their German and Russian translations (for Russian we have two translations at our disposal – see below Russian1 and Russian2).[2] It turned out that, as expected, *yes, ja* and *да*, as well as *no, nein* and *нет* function mostly as equivalents. On the other hand, in many cases these words do not coincide (about 20% for NO and about 30% for YES). Often *нет* 'no' is translated in a different way even though the respective English or German phrase with *no* resp. *nein* is acceptable. We distinguish several cases of cross-linguistic non-parallelism. Let us mention some of them.

2 Corpus Evidence

Some of these discrepancies are due to colloquial variants of YES and NO, existing in every language (e.g. *да* 'yes' – *ага* 'aha'– *угу* 'uh-huh', etc. in Russian). Cf. (1).

(1) English ***Nah**, I'm all right, thanks, Ron.*
German ***Nö**, komm schon zurecht, Ron.*
Russian1 ***Да нет**, все в порядке, спасибо, Рон.*
Russian2 ***Не-а**, все нормально, спасибочки, Рон.*

Some other phrases contain different synonyms of YES and NO, and therefore they are not parallel. Cf. (2).

(2) English *"**I don ' t think so**," said Snape, smiling coldly.*
German *„**Das kommt nicht in Frage**", sagte Snape kalt lächelnd.*
Russian1 – ***Нет, я так не думаю**, – сказал Снэйп, холодно улыбаясь.*
Russian2 – ***Ничего подобного**, – заявил Злей, холодно улыбаясь.*

Non-parallel phrases are characteristic of the situation when YES and NO are optional in some of given languages. Cf. (3).

(3) English *"Harry Potter's giving out signed photos!" "**No**, I'm not," said Harry angrily, his fists clenching.*
German *„Harry Potter verteilt Autogrammkarten" „**Nein**, tu ich nicht", sagte Harry wütend und ballte die Fäuste.*
Russian1 – *Тут Гарри Поттер раздает фотографии с автографами! – Не раздаю, – зло сказал Гарри, сжимая кулаки.*
Russian2 – *Ничего подобного, – сердито сказал Гарри, непроизвольно сжимая кулаки.*

These cases of non-parallelism are rather trivial. However, there are some fundamental grammatical differences between the words for YES and especially for NO in different languages. We will mention two of them.

The Russian *нет* 'no' can be used in contexts like *Я тоже нет* 'Neither am/do I'; *Я заплатил, а он нет* 'I paid, and he did not'; *Почему бы и нет* 'Why not'; *Разумеется, нет* 'Of course, not'*; Я думаю,*

[2] We used the Regensburg Parallel Corpus (ParaSol). Our search strategy included three iterations: we looked separately for *yes/no*, for *ja/nein*, and for *да/нет*. From this perspective, it is not so important for our aims that the original text is an English novel. We are aware of the fact that translations will not necessarily reflect pure native language intuitions. But if the translator uses a word different from the word used in the original text (a word that is not a standard equivalent), this provides even stronger evidence for the idea that there are relevant cross-linguistic differences in this field.

что нет 'I think not' etc. English *no* and German *nein* are not used in these phrases, they are replaced with *not, neither* and respectively *nicht*. So, it is a special reading of the Russian word *нет* which corresponds neither to the English *no*, nor to the German *nein*. Cf. (4-6).

(4) English *"Definitely **not**," said George, sniggering.*
German *„Ganz bestimmt **nicht**", sagte George wiehernd.*
Russian1 *– Ну конечно же, **нет**, – добавил, хихикая, Джордж.*
Russian2 *– Ни в коем случае, – хищно ухмыльнулся Джордж.*

(5) English *Scabbers has been fighting, **not** us.*
German *Krätze hat gekämpft, **nicht** wir.*
Russian1 *Скабберс дрался, мы **нет**.*
Russian2 *Это Струпик дрался, а не мы.*

(6) English *"Of course **not**," said Hermione briskly.*
German *„Natürlich **nicht**", sagte Hermine energisch.*
Russian1 *– Конечно, **нет**! – оживилась Эрмиона.*
Russian2 *– Разумеется, **нет**! – живо отозвалась Гермиона.*

The second constructive discrepancy is connected with the structure of disjunctive questions in different languages. Every language develops special means of expression in this domain, e.g. all possible kinds of tag questions in English. Cf. (7-11).

(7) English *You haven't been fighting, **have you**?*
German *Ihr habt euch nicht geschlagen, **oder**?*
Russian1 *Вы не дрались, **нет**?*
Russian2 *Вы не дрались, **нет**?*

(8) English *Harry didn't know whether to laugh **or not**.*
German *Harry wußte nicht recht, **ob** er lachen sollte.*
Russian1 *Гарри не знал, смеяться **или нет**.*
Russian2 *Но Гарри не понимал, смеяться ему **или нет**.*

(9) English *You do want to investigate Malfoy, **don't you**?*
German *Ihr wollt doch Malfoy aushorchen, **oder**?*
Russian1 *Вы хотите допросить Малфоя **или нет**?*
Russian2 *Вы хотите допросить Малфоя **или нет**?*

(10) English *I mean, it'd just go right through you, **wouldn't it**?*
German *Ich meine, es würde einfach durchfliegen, **oder**?*
Russian1 *Книга ведь просто пройдет сквозь тебя, **разве нет**?*
Russian2 *...я хочу сказать, он ведь прошел сквозь тебя, правильно?*

(11) English *You won ' t tease him, **will you**?*
German *Aber ihr zieht ihn doch damit jetzt nicht auf, **oder**?*
Russian1 *Вы же не будете дразнить его, **ведь нет**?*
Russian2 *Вы не будете его дразнить? **Нет**?*

As the examples show, also in Russian *нет* is not a mandatory part of disjunctive questions, cf., for instance, example (10) where *правильно* 'right' is used as a question tag.

Now let us briefly mention some other interesting cases of cross-linguistic contrast on the YES/NO-domain. *Yes* is used in English, when one asks someone to repeat a phrase. In German we would rather say *Wie bitte?* '≈ Pardon?', *Was?* 'What?' – and in Russian *А?* '≈Eh?', '≈What?',*Что?* 'What?', *Как?* 'How?'. YES and NO words also behave differently as end-particles in the languages in question (cf. Levontina, 2000; Левонтина, 2000). Compare (12).

(12) English *Another Weasley,* ***eh****?*
German *Noch ein Weasley,* ***nicht wahr****?*
Russian1 *Еще один Висли,* ***да****?*
Russian2 *Очередной Уэсли,* ***так****?*

Nein is often used in German to express astonishment, which is less typical of English and hardly possible in Russian. Cf. (13).

(13) English *"****Uh-oh****," said Ron, jabbing at the Invisibility Booster.*
German *„****O nein****", sagte Ron und hämmerte auf den Knopf für den Unsichtbarkeits-Servoantrieb ein.*
Russian1 *–* ***А, черт!*** *– Рон снова ткнул пальцем в кнопку невидимости.*
Russian2 *–* ***Ой****, – сказал Рон, хватаясь за исчезатель.*

It is obvious that in this field there is a lot of cases requiring futher detailed research.

3 Some Reasons for the Observed Cross-Linguistic Differences

In this paper it is not feasible to analyze all relevant cross-linguistic differences and their semantic foundations in detail. Howeverm we wish to discuss a number of them.

One difference between the three languages under consideration is quite evident. We have already mentioned it in connection with the Obama example. These languages express the idea of disagreement in basically different ways. The YES-NO-system is binomial in English and Russian (*да/yes* and *нет/no*), and trinomial in German: besides *ja* and *nein*, there is a word *doch*.[3]

Interestingly, the affirmative and negative words are distributed differently in English and Russian. The English *yes* can (differently from the Russian *да* 'yes') be used not only for confirmation, but also for denial. Cf. the example we began with (*Yes, we can*). Hence, the Russian equivalent of German *doch* would be *нет* 'no', and the English one would be *yes*.

Let us consider a very apt example. In "Winnie the Pooh", Piglet insisted that the inscription on his broken door plate *Trespassers W* bore the name of his grandfather. The English text goes on like this:

(14) English *Christopher Robin said you couldn't be called Trespassers W, and Piglet said* ***yes, you could****, because his grandfather was*.

Compare the Russian version by Boris Zakhoder (15).

(15) Russian *Кристофер Робин сказал, что не может быть такого имени – Посторонним В., а Пятачок ответил, что* ***нет, может, нет, может,*** *потому что дедушку же так звали!*

Naturally enough, the German translation reads:

(16) German *Christopher Robin sagte, man könne nicht Betreten V heißen, und Ferkel sagte,* ***doch, das könne man****, sein Großvater habe ja so geheißen*.

[3] In this regard, French is similar to German: The French word *si* is used within the YES-NO-system in the same way as the German *doch*.

The analysis of our data shows that in contexts of this type *yes* systematically correlates with *нет, no* with *да*. Compare the following examples.

(17) English *"**No**, sir, nor with me, sir," said Mr Borgin, with a deep bow.*
German *„**Nein**, Sir, bei mir auch nicht", sagte Mr Borgin mit einer tiefen Verbeugung.*
Russian1 *– **Да**, сэр, и не для меня, сэр, – подтвердил мистер Борджин с глубоким поклоном.*
Russian2 *– **Разумеется**, сэр, и не я тоже, – с глубоким поклоном подтвердил мистер Борджин.*

(18) English *"**No**," said Harry, getting into bed.*
German *„**Nein**", sagte Harry und stieg ins Bett.*
Russian1 *– **Да**, – ответил Гарри, забираясь в постель*
Russian2 *– Не болит, – согласился Гарри, забираясь в постель.*

(19) English *"This is a girls' bathroom," she said, eyeing Ron and Harry suspiciously. " They're not girls." "**No**," Hermione agreed. "I just wanted to show them how er – nice it is in here."*
German *„Das ist ein Mädchenklo", sagte sie und musterte Ron und Harry mißtrauisch. „Das sind keine Mädchen." „**Nein**", stimmte ihr Hermine zu, „ich wollte ihnen nur zeigen, wie – ähm – nett du es hier hast."*
Russian1 *– Этот туалет для девочек, – сказала она, подозрительно оглядывая Рона и Гарри. – А они – не девочки. – **Это точно**, – согласилась Эрмиона. – Я только хотела им показать, как здесь мило.*
Russian2 *– Это туалет для девочек, – заявила она, с подозрением воззрившись на мальчиков. – А они не девочки. – Не девочки, – согласилась Гермиона. – Я просто привела их взглянуть, как тут... интересно.*

As these examples show, the Russian word *нет* shows profound differences from its English and German equivalents; cf. (Добровольский & Левонтина, 2009). These differences can be described in terms of scope; cf. (Богуславский 1996). Broadly, we can consider the English words *yes* and *no* to be oriented, above all, towards the inner scope, i.e. towards the propositional content of the utterance. On the contrary, the Russian words *да* 'yes' and *нет* 'no' are rather oriented towards the outer scope, i.e. they express agreement or disagreement with what has been said or even meant in previous utterances. This issue is connected not only with the meanings of these very words, but also with the dialogue strategies favoured by given languages. In English, we can say *You cannot do it – Of course, I can*. However, in Russian the dialogue: *Вы этого не можете – Конечно, могу* would sound incoherent. This example shows that not only the semantics of *yes* or *da* 'yes' is decisive, but, above all, the fact that a Russian speaker has to explicitly mark that he or she disagrees with the interlocutor by using words such as *нет* 'no' or *a vot i* '≈ and still', *esche kak* '≈ and still' etc. To put it another way, the dialogue strategy of expressing disagreement is lexicalized in Russian. Compare also examples (14-19) above.

4 Does the Russian *нет* have special additional readings?

If the problem consisted only of examples, such as those discussed here, it could easily be solved by a simple rule of syntactic nature. However, there are many other instances of non-coincidence between the usage of *нет* 'no' and its English and German equivalents. Such examples may look not quite convincing in every single case, but, in sum, they prove that here we are dealing with fundamental semantic differences.

One interesting group of contexts where we find clear differences between English and German, on the one hand, and Russian, on the other, consists of sentences which contain, so to speak, "reactions to the

unexpressed". Compare examples (20-22) where authentic Russian sentences are presented with their literal translations which we produced as a linguistic experiment.

(20) Russian ***Нет**, а что ты лезешь со своими советами*!
German ****Nein***, *was kommst du mit deinen Ratschlägen!*
English ****No***, *what do you come with your recommendations?*

(21) Russian ***Нет**, а вам какое дело!*
German ****Nein***, *was geht Sie das an!*
English ****No***, *is it your business?*

(22) Russian – *Вы опоздали. И юбка у вас слишком коротка. – **Нет**, а юбка-то здесь причем?*
German *„Sie kommen zu spät! Und Ihr Rock ist definitiv zu kurz." „**Nein**, was hat denn mein Rock damit zu tun?"*
English *"You come too late. And your skirt is definitely too short." "**No**, what's wrong about my skirt?"*

Here is a very apt example from N.V. Gogol's "Dead Souls", taken from the Russian-English parallel corpus developed as a component of the Russian National Corpus (RNC):

(23) Russian – *Извинительней сходить в какое-нибудь непристойное место, чем к нему. – **Нет**, я спросил не для каких-либо, а потому только, что интересуюсь познанием всякого рода мест, – отвечал на это Чичиков.*
English *"A man had far better go to hell than to Plushkin's." "**Quite so**," responded Chichikov, "my only reason for asking you is that it interests me to become acquainted with any and every sort of locality."*

The translation of *нет* 'no' by *quite so* is not an accidental translator decision, but a reflection of systematic cross-linguistic differences in the semantic organisation of the YES-NO-domain. It is not possible to use *no* in English if the speaker basically agrees with the interlocutor. The Russian *нет* 'no' here denotes the denial of the unexpressed idea that Chichikov tries to derive benefit from visiting Plushkin. Thus we are dealing here with a "reaction to the unexpressed".

A further non-trivial example (24) that we encountered in the "Dead Souls" refers to a situation in which the speakers use *нет* 'no' as an indicator of their own, not mentioned so far, hesitation or vacillation.

(24) Russian *Письмо начиналось очень решительно, именно так: «**Нет**, я должна к тебе писать!»*
English *Beginning abruptly with the words "I MUST write to you!" the letter went on to say that <...>*

In this case, neither the English *no* nor the German *nein* may be used.

Paradoxically, the Russian negation *нет* 'no' is often used in dialogue to express the full agreement with the interlocutor. The most suitable explanation goes like this: The speaker agrees with the interlocutor so eagerly that he/she wants to immediately take the initiative. This negation-word can be used in Russian colloquial speech with the function of turn-taking. Here are some examples of this special reading of *нет* (the same as in (20-22), Russian sentences were observed in real dialogues and then translated into German and English).

(25) Russian – *Давай напишем статью вместе! – Нет, я тоже уже об этом думал.*

German – *Schreiben wir einen Artikel zusammen! – *Nein, ich habe auch dran gedacht*.
English – *Let us write a paper together. – *No, me too, I also thought about that.*

(26) Russian – *Мы должны позвонить Пете. – Нет, правильно. А то он опоздает.*
German – *Wir müssen noch Peter anrufen. – *Nein, richtig. Sonst kommt er zu spät.*
English – *We have to give a ring to Pete. – *No, right. Otherwise he will be too late.*

Combinations such as *нет, правда* 'no, right'; *нет, точно* 'no, exactly'; *нет, конечно* 'no, of course'; and even *нет, да* 'no, yes' (cf. *Нет, да, я согласен* 'No, yes, I agree') are very typical of Russian discourse, whereas combinations like German *nein, stimmt* 'no, right'; *nein, ja* 'no, yes' or English *no, sure; no, yes* would violate the usage norms of English and German. Obviously, the reason for this is that *no* or *nein* has no phatic function, whereas this function is typical of the Russian *нет*.

In all instances of this kind, the Russian word *нет* 'no' interacts with the speaker's ideas about the situation as a whole, rather than with the proposition. In other words, what is negated is not some already mentioned issues, but "the unexpressed". Saying, e.g., *Нет, это уж слишком!* 'No, it is too much!', we do not give a negative answer to some question (nobody has asked us anything), but express our rejection of the discussed situation. In a similar way, utterances such as *Нет, ты представь себе!* 'No, imagine!' sound in many situations more natural than *Ты представь себе!* 'Imagine!'.[4] The speaker quasi objects to the very possibility that he or she meets with no sympathy from the addressee. So, here we are obviously dealing with a special, phatic reading of *нет*.

In our analysis of parallel corpora, we regard cases when the translator adds a word which is absent in the original text as being very significant. Obviously, good translators make use of such modifications only if they feel that the target-language utterance in question would sound not natural enough or show some lack of coherence without this additional word. Cf. (27-29).

(27) English *Harry backed away. "I'm fine, thanks," he said.*
German *Harry wich zurück . „Geht mir gut, danke", sagte er.*
Russian1 *Гарри отпрянул. – **Нет-нет**, – сказал он, – все в порядке.*
Russian2 *Гарри попятился. – Со мной все в порядке, спасибо, – ответил он.*

(28) English *"He's in Gryffindor," said Harry quickly.*
German *„Er ist ein Gryffindor", sagte Harry rasch.*
Russian1 – ***Нет**, он из «Гриффиндора», – поспешил разубедить его Гарри.*
Russian2 – *Он из «Гриффиндора», – быстро сказал Гарри.*

(29) English *I don't believe it!*
German *Ist doch nicht zu fassen!*
Russian1 ***Нет**, вы подумайте!*
Russian2 *Не могу в это поверить!*

There is one more reading of the Russian *нет* 'no' that has no exact correlates in English and German, namely its use in imperative sentences. These cross-linguistic differences are especially obvious in the situation of repeated inducement; cf. (Левонтина, 1999). In case of an explicit refusal, the repeated demand often begins with *нет*, and it is normally inadmissible. Cf. the following example from the German-Russian parallel corpus (RNC).

[4] Utterances such as *No, it is too much!* or *No, imagine!* are possible in English, but the function of *no* is here different from the function of *нет* in the corresponding Russian sentences. *No* bears a stress and is interpreted here as the expression of astonishment and not as a phatic signal.

(30) German *„Kommen Sie mit!“ wiederholte K. jetzt schärfer, als habe er endlich den Gerichtsdiener auf einer Unwahrheit ertappt.*
Russian – ***Нет*** *пойдемте! – уже резче сказал К., словно наконец уличил служителя во лжи.*
literally ‘No, come with me!” said K. again, more sharply, as if he finally caught the officer in a lie.’
(Franz Kafka. Der Prozeß)

For more detail see (Добровольский & Левонтина, в печати).

5 Conclusion

Many of the cross-linguistic differences discussed above relate to the specific properties of the discourse structure. Thus, the Russian discourse is characterized by a hypertrophied coherence, addressing various levels of content, and focusing on interpersonal relationships between interlocutors, which they steadily profile in the course of conversation.

References

Levontina Irina B. 2000. Abschlusspartikeln im Russischen und Deutschen. In F. Maurice & I. Mendoza (eds.): *Linguistische Beiträge zur Slavistik VIII.* 145-153. Sagner, München.

Богуславский Игорь М. 1996. *Сфера действия лексических единиц*. М.: Школа «Языки русской культуры».

Добровольский Дмитрий О. & Ирина Б. Левонтина. В печати. 500 способов сказать «нет» (русско-немецкие соответствия) // *Логический анализ языка. Ассерция и негация*. М.: Индрик.

Добровольский Дмитрий О. & Ирина Б. Левонтина. 2009. Русское *нет*, немецкое *nein*, английское *no*: сопоставительное исследование семантики на базе параллельных корпусов // *Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» =Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue”.* Выпуск 8 (15). М.: РГГУ.

Левонтина И рина Б. 1999. Стратегии уговаривания: частицы в повторных просьбах // *Язык. Культура, Гуманитарное знание. Научное наследие Г. О. Винокура и современность.* 188-201. М.: Научный мир.

Левонтина Ирина Б. 2000. Русское финальное *а?*: Портрет невидимки // *Слово в тексте и в словаре. Сборник статей к семидесятилетию академика Ю.Д. Апресяна*. М.: Языки русской культуры.

French infinitive in Arabic translation: a usage-based approach in MTT

Dina El Kassas
French department, Al Alsun Faculty
Minya University, Egypt
delkassas@gmail.com

Abstract

This article deals with the Arabic syntactic counterparts of the French infinitive in various syntactic contexts. Special attention is paid to infinitive occurring as head of independent clause and expressing injunctive modality. Syntactic transfer is examined in order to show the necessity of a usage-based approach, which avoids giving equivalents that are grammatically correct, but weird according to common use.

1 Introduction

The infinitive is a component of many commonly used French grammatical constructions. Its complex grammatical nature is due to its high polyfunctionality. The infinitive form of the French verb may function as verb, noun, adjective or adverb (section 2).

In the present paper, we are exploring the Arabic structural equivalents of the French infinitive. This equivalent is not always a non-finite verb form, and therefore a decision must be made about the Arabic verb's tense as well as agreement in person, gender and number (PNG agreement). In some cases, choices may be made by resolving anaphoric reference within the sentence or in the preceding text (section 3). But in other cases, the choice is oriented by common usage in the target language (here, the Arabic); and we have to define the kind of text where the infinitive occurs in order to choose the appropriate counterpart. As an example, we will examine the Arabic equivalents of the infinitive expressing injunctive modality and heading a main clause through different contexts (section 4).

2 French infinitive

The infinitive is defined as a type of non-finite verb allowing the expression of a process in its most virtual form, or the action as a general concept, without specifying the circumstances, given that the form of the verb doesn't change: the infinitive form of the verb is not marked by categories such as tense, person, gender or number. Thus, it doesn't agree with a subject — the one responsible for the action in case of an infinitive is "everybody"; otherwise, the person concerned by the action in an infinitive is announced in the sentence. The abstract nature of the act evoked by an infinitive explains its usage as an entry in dictionaries. That's why we often encounter it in sentences that say something general, like *Tricher n'est pas jouer* 'Cheating is not playing' or *Il est important de faire ses devoirs* 'It is important to do homework'.

The infinitive is classified as a grammeme (= inflectional meaning) of finiteness, the inflectional category of the verb as syntactic head. [1] The finiteness category counts five grammemes: finite, infinitive, maṣdar, participle and gerundive. All languages don't necessarily include these five values.

[1] Mel'čuk, Igor A. 1994. *Cours de morphologie générale*, volume 2. University Press of Montréal, Canada, p. 215.

French counts three infinitive forms: simple, compound and double-compound forms. Therefore, infinitive may express relative time, as simultaneity or successivity. Rémi-Giraud (1988) distinguishes two values of the infinitive:

1. Present infinitive in opposition with past infinitive
2. Present infinitive form not in opposition with a past form

The infinitive form of a verb functions as a verb, a noun, an adjective or an adverb.

As verb, infinitive may head a main clause:

- narrative infinitive: *Et Paul de rajouter* 'Then, Paul started adding things'.
- deliberative infinitive: *Où aller ?* 'Where to go?'.
- exclamative infinitive: *Quoi ! ne pas dire à son patron !* 'What! Not telling his boss'.
- imperative infinitive: *Battre les œufs en neige* 'Stiffly beat the egg whites'.

The Arabic counterpart in such cases is generally an inflected verb. In order to give the appropriate translation in Arabic, we have to examine temporal or modal value of the infinitive verb and examine situational context to check for the structure commonly use, cf. section (4).

On the other hand, following the Latin model, the infinitive may be embedded within multiple syntactic structures. The first syntactic actant of the main verb is not always co-referent with the subject of the embedded infinitive verb.

The embedded infinitive can be governed by:

- a perception verb: *J'entends les enfants crier* 'I hear children yelling',
- a movement verbs: *Il a couru chercher son livre* 'He ran to search his book',
- a modal or a light verb: *Je dois fermer la porte* 'I must shut the door', *Il commence à parler* 'He starts speaking',
- the verb *aller* in a construction expressing future: *Je vais partir* 'I will go',
- the verb *venir* in a construction expressing recent past: *Je viens de sortir*, 'I just came out',
- a speech verb in interrogative indirect speech: *Je ne sais plus où aller* 'I don't know where to go',
- a relative pronoun: *Elle cherche une sale où fêter son anniversaire* 'She is searching for a place where to celebrate her birthday'.

The French infinitive can also perform:

- Nominal functions, such as subject (*Travailler est une nécessité*, 'Working is a necessity'), object complement (*Il préfère partir*, 'He prefers leaving' / *Il incite Jean à se reposer*, 'He encourages John to take a rest') and noun complement (*La peur de courir des risques*, 'The fear of taking risks'),
- Adjectival functions, such as objectival attributive (*J'appelle cela tricher*, 'I call it cheating') and complement of adjective (*Un paysage agréable à regarder*, 'a view pleasing to the eyes'),
- Adverbial functions, such as circumstantial (*Il faut manger pour vivre*, 'You must eat to live' / *Avant de dormer, tu feras tes devoirs*, 'You will do your homework before going to bed'),
- In apposition (*Il n'a qu'un souhait: voyager*, 'He has just one wish: travelling').

The complexity and the profusion of the French infinitive cannot be embraced within an article, that's why we ignored for example infinitive chains like *J'ai fait promettre à Jean de partir* 'I made John promise to go'.

3 Translating the particularities of the French infinitive into Arabic

Arabic doesn't have an infinitive form. Instead, as a non-finite form of the verb, it counts a verbal noun, the maṣdar. Therefore, maṣdar is the equivalent by default of the infinitive. In this case, the translation makes no problem as the assertive modality induced by the infinitive is rendered by an equivalent uninflected verbal form, so no decision has to be made concerning person, gender and number agreement to choose the right flectional form.

In this section, we will list Arabic counterparts of the French infinitive according to its syntactic context. We will present the following equivalents: the maṣdar, the completive clause introduced by *ʔan*, the completive clause introduced by *ʔanna*, the adjectival equivalent, the status clause and the finite verb.

3.1 Translated by a maṣdar ($V_{inf} \equiv N_{Maṣdar}$)

The maṣdar presents the commonly used infinitive equivalent, wherever the infinitive functions

– as a verb heading a main clause (1a) or as a subordinated complement (1b-c):

(1) a. A quoi bon **preparer** cet examen ? 'What's good in preparing this exam?'

maa	faaʔidatu	**al–taḥḍiiri**[2]		li=haðaa	al–ʔimtiḥaani
what	utility	the+preparation:Maṣdar		for=this	the–exam

b. Elle doit **respecter** la loi 'She must respect the law'

jaʤibu	ʕalaj=haa	**ʔiḥtiraamu**	al–qaanuuni
must.V	on=her	respecting:Maṣdar	the–law

c. Il a passé[3] des heures à **jouer** 'He spent hours playing'

ʔamḍaa	$\varnothing_{\text{pronoun:subject}}$	saaʕaati–n	fii **al–laʕibi**
he spent		hours	in the–playing:Maṣdar

– as a noun, fulfilling functions such as subject (2a), quasi subject (2b) and direct object (2c):

(2) a. **Reussir** est mon but 'My goal is succeeding'

al–naʤaaḥu	hadaf–ii[4]
the–success:Maṣdar	target–my

b. Cela fatigue Alain **de courir**[5] 'Running is exhausting Alain'

jurhiqu	**al–ʤarju**	ʔalaan
is exhausting	the–running:Maṣdar	Alain

c. Je ne sais pas **nager** 'I don't know how to swim'

laa	ʔaʕrifu	$\varnothing_{\text{pronoun:subject}}$	**al–sibaaḥata**[6]
Neg.	I know		the–swimming:Maṣdar

– as an adjective, fulfilling functions such as noun or adjective complement:

(3) a. La fiche à identifier[7] 'The form to identify'

al–ʔistimaaratu	al–waaʤibu	maʕrifatu=haa
the–form	the–obliged_to	identifying:Maṣdar=it:PRO

---

[2] The deliberative infinitive in this example may also be translated by a completive clause, ʔan + Verb in the subjunctive marked with the 2sd singular masculine person: *maa faaʔidatu* ***ʔan tuḥaḍira*** *lihaðaa alʔimtiḥaani*. But as we will see, this is not a systematic transfer: the verb is not always in the subjunctive mood and doesn't systematically agree with 2sd singular masculine person. PNG agreement may differ according to the speaker (the addressee by the direct discourse), the situational context and the usage.

[3] This type of trivalent verbs accepts alternatively as equivalent a maṣdar or a verb in the present tense expressing progressivity. Therefore, the mentioned example can be translated as *ʔamḍaa saaʕaatin* ***jalʕabu***. The verb is marked by a pronominal subject co-referent with the subject of the main verb.

[4] Again switching to a completive clause is possible but the verb has to be inflected in the 1st singular person: *hadaf–ii* ***ʔan ʔanaʤaaḥa***. However, this commutation is not always possible, sometimes translating by a completive clause is not possible: **De voir** ça m'a bouleversé ≡ *ʔinzaʕaʤtu* ***min ruʔjati*** *haðaa* ('I was upset by seeing that') | **ʔinzaʕaʤtu* ***ʔan ʔaraa*** *haðaa*.

[5] Several impersonal constructions conducting general values or expressing conative function (like advices, orders or prayers) accept as equivalent indiscriminately a masdar or a completive clause. The verb of the complement is marked with the 2sd masculine singular person or the 1st plural person:

Il est préférable **d'être** poli ≡ *mina almufaḍali* ***ʔan takuuna muhaðaban*** | ***altahðiibu*** *mufaḍalun* 'It is preferable to be polite'

Il est dommage **de fumer** autant ≡ *mina almuḥzini* ***altadχiinu*** / ***ʔan tudaχina*** *bihaðihi alʃaraahati* 'It is sad smoking that much'

[6] In this example, the commutation with a completive clause is not possible, yet the infinitive may be translated by a verb in the present tense inflected with a person pronoun co-referent with the subject of the main verb, here, the 1st singular person: *laa ʔaʕrifu kajfa* ***ʔasbaḥu***.

[7] We will not discuss the translation of the preposition *à* according to its numerous uses.

b. Cet exercice est facile **à faire** 'This exercise is easy to do'
haðaa al–tamriinu $\varnothing_{kaana}$ sahlun **ḥalu=hu**
that the–exercice is easy doing:Maṣdar=it:PRO

Although the infinitive can be translated by a maṣdar, Arabic shows divergences in syntactic structure as well as word order constraints due to the particularity of the stative sentence governed by the copula *kaana*. The syntactic structure may be presented as follow:

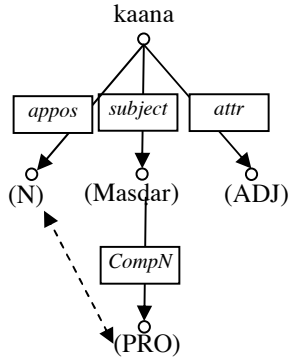


Figure 1**: Surface Syntactic Structure of an Arabic copular clause**

– as an adverb, fulfilling functions such as a circumstantial with a preposition:

(4) a. Tu parles sans **reflechir** 'You speak without thinking'
ʔinna=ka tataḥadaθu $\varnothing_{pronoun:subject}$ duuna **tafkiirin**
Assert=you you speak without thinking:Maṣdar

b. Sans entrer dans des détails, ne faites plus cela 'Without going through details, don't do that again'
duuna al–duχuuli fii al–tafaaṣiili, laa taʕuud li=maa faʕalta
without entering:Maṣdar in the–details Neg you return to you do

The fact that the maṣdar is the default equivalent of the infinitive doesn't mean that the same syntactic relations occur in both languages. These relations are language-specific and divergences in syntactic patterns are frequent. For example, the French quasi-subject can be promoted to be a subject in Arabic:

(5) *Cela fatigue Alain **de courir*** α jurhiqu **alʤarju** ʔalaan ('Running is exhausting Alan')

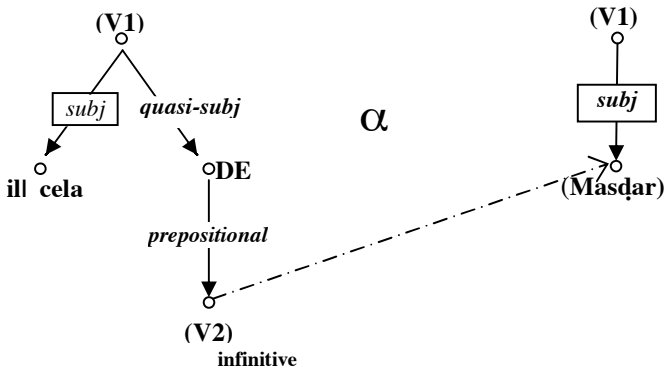


Figure 2: quasi-subject promoted to a subject in the Arabic structure

3.2 Translated by a completive clause introduced by ʔan

(V_{inf} ≡ ʔan–conjunctive→$V_{subjunctive}$–subject→$PRO_{co\text{-}reference\ with\ main\ clause}$)

Translating an infinitive by the homologue grammeme, the maṣdar, is not always appropriate, even though it may be grammatically correct. If in a context implying general injunctive modality, the maṣdar can be successfully used as the counterpart of the infinitive, in case of a context implying injunctive value in a specific situation, the completive clause introduced by the complementizer *ʔan* presents the equivalent commonly used. The subordinated verb is in the subjunctive and takes a pronominal subject. The choice of the appropriate pronoun is not obvious because the first syntactic actant of the infinitive verb is not always co-referent with the subject of the main verb; for example, in (6b) it is co-referent with the attribute and in (6c) with the direct object.

(6) a. Elle doit **partir** α jaʤibu ʕalaj=haa **ʔan tarḥala** $\varnothing_{PRO:subject}$ 'She must go'

b. C'est pour moi un moment émouvant que **de me retrouver** à l'université α ʔinna=haa la=laḥẓatun muʔaθiratun lii **ʔan ʔatawaaʤadu** $\varnothing_{pronoun:subject}$ bi=l–ʤaamiʕati 'It is an emotional moment to me to be at the university'

c. Se voir sortir m'irrite α juzʕiʤunii **ʔan ʔaʤida** $\varnothing_{pronoun:subject}$ nafsii χaariʤatun 'It is irritating me to be outside'

3.3 Translated by a completive clause introduced by ʔanna

The embedded infinitive governed by a speech or perception verb can be translated also by a finite verb. The Arabic main verb must occur with the complementizer *ʔanna*. The fact that in French the 1st syntactic actant of the main verb and that of the infinitive have the same referent resolves the PNG subject agreement, and the choice of tense is fixed as follows: past infinitive α past tense, present infinitive α present tense, cf. (7).

(7) Ces poires paraissent avoir voyagé 'the pears seems to have travelled'

jabduu	**ʔanna**	**haðihi**	**al–kumeθra**	**saafarat** $\varnothing_{pronoun:subject}$
seems	that	this	the–pears	it has travelled

3.4 Translated by a relative clause as adjectival equivalent

An infinitive modifying a noun can be translated by a relative clause (8). It is mainly the case of the prepositional phrase "à + V_{inf}". The transferring rule is as follow: {N–**Modification**→à–**Prepositional**→ V_{inf}} ≡ {N–**Modification**→Relative:Connector–**Conjunction**→$V_{present}$}. The Arabic subordinated verb is in the present tense and takes a subject co-referent with the heading noun:

(8) Vous êtes les premiers à le voir 'You are the first to see it'

ʔantum	ʔawalu	**man**	**juʃaahida=hu**	$\varnothing_{pronoun:subject}$
You	first	who	you see=it	

Figure 3 shows the correspondence between the French structure and its Arabic equivalent:

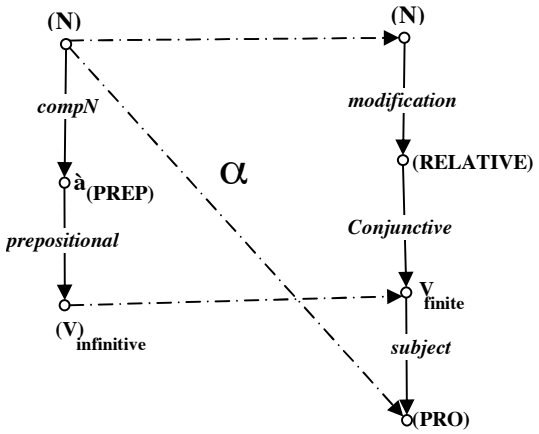


Figure 3 : The French infinitive translated into Arabic by a relative clause

3.5 Translated by a status clause

The embedded infinitive, in particular the *infinitivus cum Accusativo* construction, cf. (9a),[8] may also be translated by a status clause (a coordinate clause introduced by the conjunction wa 'and' and acting as a modifier). The status clause has the specifity of being followed by a finite verb in the present tense and agreeing with a pronominal subject co-referent with the noun modified. This noun can be the subject of the infinitive clause (9a) or correspond to the first syntactic actant of the infinitive phrase fulfilling a circumstantial function (9b).

(9) a. Nous avons regardé **la scène se derouler** 'We watched the scene unfolding'

ʃaahadnaa ∅$_{\text{pronoun:subject}}$	al–ʔaḥdaaθu	**wa hija**	**taduuru'**
we watched	the–events	and it	it is unfolding

b. Le criminel a tué Jean **sans laisser** de traces 'The criminal killed John and didn't left evidence'

qatala ∅$_{\text{pronoun:subject}}$	al–mudʒrimu	dʒon	**wa lam**	**jatruka**	**ʔaθarun**
he killed	the–criminel	John	and negation	he left	evidence

3.6 Translated by a finite verb

Translating the French infinitive into Arabic by a finite verb requires decisions concerning the tense and its agreement with a subject. This indeterminacy may occasionally entail resorting to linguistic or situational context. The Arabic counterpart can be a verb in the present, past or future tense. It can be also in the imperative or the passive. In this subsection, we will examine the different Arabic finite verb forms functioning as counterparts of the French infinitive.

– **Translating the French infinitive by a verb in the present tense**

The embedded infinitive can be translated by a verb in the present tense, when governed by a main verb expressing progressivity, cf. (1c). The Arabic pattern may present a paratactic construction, {$V_{finite} \rightarrow V_{finite}$}, when governed by a verb of perception (10a), an inchoative or a continuative verb heading as oblique infinitival object[9] the prepositional phrase "de + $V_{infinitive}$" or "à + $V_{infinitive}$" (10b). In

[8] The *infinitives cum assusativo construction* or the *accusative plus infinitive* construction (A+I) involves a verb (in (9a) is verb *regarder*) followed by a noun phrase (la scène) and an infinitive (se dérouler). It can be rephrased by a sentence containing a complement clause (Nous avons regarder la scène **qui se deroule**).

[9] In French, the infinitival oblique-objective syntactic relation is to be distinguished from the oblique-objective one, cf. (Iordanskaja & Mel'čuk, 2000). Correspondingly, to describe the paratactic relation between the main verb and the embedded

these cases the corresponding Arabic construction must be given in the government pattern of the verb. But the paratactic construction may also express a circumstantial value, for instance with a verb of movement as a main verb, cf. (10c). In the three cases, the 1st syntactic actant of the infinitive verb is co-referent either with the subject or the object of the main clause. Resolving this co-referential issue will help choosing the right inflectional form of the Arabic counterpart:

(10) a. Pierre la vit tuer 'Pierre saw her killing [someone]'
raʕa=haa $\varnothing_{\text{pronoun:subject}}$ pjir taqtulu $\varnothing_{\text{pronoun:subject}}$
he saw=her Pierre she is killing

b. Les enfants ont commencé à dormir 'Children begin to sleep'
badaʔa al–ʔawlaadu janaamuun. $\varnothing_{\text{pronoun:subject}}$
start the–children they are sleeping

c. Il court acheter des cigarettes 'He runs to buy cigarettes'
ʤaraa $\varnothing_{\text{pronoun:subject}}$ jaʃtarii $\varnothing_{\text{pronoun:subject}}$ saʤaaʔirun'[10]
he runs he buys cigarettes

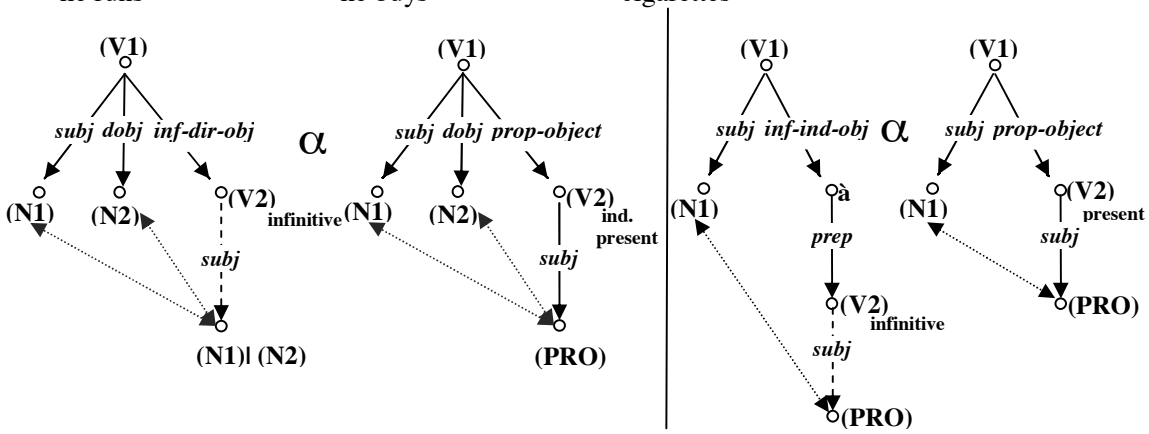


Figure 4: French infinitive translated by a verbal paratactic construction into Arabic

A verb in the present tense may also occur as the appropriate equivalent of the embedded indirect interrogative clause (11a) as well as the deliberative infinitive governing independent clause (11b):

(11) a. Je ne sais plus où aller. 'I don't know where to go'
laa ʕadrii $\varnothing_{\text{pronoun:subject}}$ ʕajna ʕaðhabu $\varnothing_{\text{pronoun:subject}}$
Neg I know where I go

b. Que faire ? 'What to do?'
maaðaa ʔafʕalu $\varnothing_{\text{pronoun:subject}}$
what I do

In (11a), the co-referential link between the 1st syntactic actant of the main verb and the infinitive resolves the problem of choosing the pronominal subject in the corresponding subordinate clause. In (11b), the choice of the 1st masculine singular pronoun as the subject is justified since the speaker is the enunciator or the lecturer himself. But the decision is not always obvious: for instance, with an

finite one in Arabic, we distinguish a propositional objective relation. The propositional object may occasionally be substituted by a maṣdar or a preposition + maṣdar. As a result, we can say *badaʔa alʔawlaadu* ***alnawma*** or *aχaða alʔawlaadu* ***fii alnawmi***.

[10] The subordinated clause may be translated by a circumstantial clause introduced by the subordinating conjunction *li* 'in order to'. Thus, we can say *ʤaraa* ***li=jaʃtarii*** *saʤaaʔirun* and *ðahaba* ***li=juqabilu*** *ʕamata=hu*.

impersonal structure, we should know the addressee in order to choose the appropriate pronominal subject of the Arabic counterpart. We should know if whether it is a general injunction or it is addressed to a specific person. Thus, in (12), the embedded elements (the verb and the pronoun headed by the maṣdar) agree with the 1st plural person, even though nothing in the French sentence justifies such a choice:

(12) Il faudrait **savoir** où **aller** 'We should know where to go'
janbaġii ʔan naʕrifu $\varnothing_{\textbf{pronoun:subject}}$ widʒhatu=naa
must that we know direction=our:Maṣdar

In an independent clause, the 1st syntactic actant of the infinitive may be realized not as a subject. For example, in the exclamative sentence (13), it is realized as an apposition:

(13) Toi, voter pour lui ! 'You vote for him!'
ʔanta tantaχiba=hu $\varnothing_{\textbf{pronoun:subject}}$!
you you vote =him

The choice can be more complicated. As example, what can justify that the maṣdar in (14a) and the present tense in (14b) are commonly used rather than the opposite.

(14) a. Où s'enfuir? 'where to flee'
ʔajna al−mafarru? | ??ajna ʔafirrulnafirru?
where the−escape:Maṣdar where I escapel we escape

b. Où aller? 'Where to go?'
ʔila ʔajna ʔaðhabu l naðhabu? | ??ila ʔajna al−ðihaabu?
to where I go l we go to where the−going:Maṣdar

The Arabic counterpart may present a fusion of the construction $\{V_{finite} \rightarrow V_{infinitive}\}$ in one finite verb. It is frequently the case of the causative construction $\{faire \rightarrow V_{infinitif}\}$, where the counterpart is either $\{dʒaʕala$ 'make' $\rightarrow V_{present}\}$ or a lexical verb derived from the causative pattern, the third verbal morphological pattern if the verb has a causative form. The choice depends on Arabic lexis, for example constructions such as *faire parler*, *faire manger* or *faire crier* will be translated by jadʒʕala=hu jataḥadaθu 'make him speek', jadʒʕala=hu jaʔkulu 'make him eat', *jadʒʕala=hu jaṣruχu* 'make him scream' while other constructions will be translated by one single verb, such as *faire dormer*, ʔanaama 'make him sleep', *faire lire*, ʔaqraʔa 'make him read', *faire sortir*, ʔaχradʒa 'put him out', *faire comprendre*, ʔafhama 'make him understand'. We will not discuss in this paper the syntactic transfer of French causative construction. The numerous syntactic restructurings of this construction can be the subject of a further research.

– **Translating the French infinitive by a verb in the past tense**

The French infinitive can be occasionally translated by a verb in the past tense, as example the exclamative infinitive in (13) can be translated by a verb in the past tense: *ʔanta **ʔintaχabta=hu** l **ṣawatta la=hu***. The subject of the Arabic verb is also a pronoun co-referent with the noun in apposition.

– **Translating the French infinitive by a verb in the future tense**

It is the case of the construction $(V_{finite} \rightarrow Aller_{infinitive})$ expressing future:

(15) Regarde, Maman, je **vais plonger**! ɑ ʔunẓurii yaa ʔumii **sawfa aġṭusu** 'look Mother, I will dive'

– **Translating the French infinitive by a verb in the passive**

By examining parallel corpora, like cooking recipes and instructions in case of natural disasters, we found that the infinitive verb can be translated by an Arabic verb in the passive. We noticed that the infinitive verb can govern an independent clause. In this case, the direct object of the infinitive verb is

promoted subject in the Arabic counterpart, cf. (16a) and (16b).

(16) a. **Cuire** la pâte lit. 'The dough is baked'
tuslaqu *–subj→* al–makaruunatu
is baked the–dough

b. **Faire revenir** l'oignon dans l'huile lit. 'The onion is browned in oil'
juḥammaru *–subj→* al–baṣalu
is browned in oil the–onion

– **Translating the French infinitive by a verb in the imperative**

The injunctive infinitive may also be translated by a verb in the imperative form. Thus, the above examples may be translated also by: *ʔusluqi al–makaruunata* and *ḥammirii al–baṣala*. The choice of the right inflectional form is guided by common usage. Therefore, in a cooking recipe, the imperative verb is inflected in the 2sd feminine singular person. In other instructional contexts, like traffic code or manuals, the equivalent is a verb in the imperative marked by the 2sd masculine singular person:

(17) a. Ralentir α χafif al–surʕata 'Slow down the car'

b. Appuyer, agiter α ʔiḍġaṭ, ruʤa 'Press, shake'

In the above contexts the translation by a maṣdar is rejected even if it's grammatically correct: ?***taχfiifu*** *al–surʕati*, ?***al–ḍaġṭu, al–raʤʤu***. In the next section, we will examine in more detail the infinitive governing an independent clause and fulfilling a conative function. We will see how the common use rejects translations that are grammatically corrects.

4 Translating the French infinitive expressing an injunctive value

In French, as the context requires, a verb carrying injunctive value may take three forms: the instructions can be in the imperative (***Entourez*** *les adjectifs invariables en rouge* 'Surround invariable adjectives in red'), in the infinitive (***Entourer*** *les adjectifs invariables en rouge*) or in the future tense (***Vous entourerez*** *les adjectifs invariables en rouge*). The infinitive can fulfill so a conative function, oriented toward the addressee. It expresses imperatives, directives and apostrophes such as orders, warnings and prayers. This infinitive is encountered in instructions of several types, e.g. manuals, recipes, advertisements, advices, administrative formalities and propaganda. It is used to assert that an action must be performed.

On the other hand, by exploring parallel corpora – mainly extracted from Egyptian and French newspapers and magazines, we see that Arabic tends to clarify the pragmatic modality of the infinitive by using a verbal form other than the maṣdar, the default non-finite counterpart. The assertive modality in constructions where French systematically uses the infinitive is also rendered in Arabic by constructions other than the maṣdar. The choice is based on standard use: a translation can be rejected even if it is grammatically correct. In such situations, the choice of a correct syntactic equivalent depends on the context For example, the following infinitives require different equivalents:

1)	Muscler le ventre	maṣdar
2)	Modifier mes données personnels	maṣdar
3)	Acheter du lait	maṣdar
4)	Eplucher les pommes	passive, imperative 1 pl, imperative 2 sg/fem
5)	Vérifier l'eau et l'électricité	maṣdar, imperative 2 sg/masc
6)	En cas de fuite, ouvrir les fenêtres	maṣdar, imperative 2 sg/masc
7)	Avoir une bonne conduite	completive clause introduced by *ʔan*
8)	Remplir les conditions	completive clause introduced by *ʔan*

In all these examples, the infinitive expresses an injunctive modality within different types of texts. The choice of the appropriate Arabic equivalent is guided by the nature of the text itself. In the first example, the text is about instruction to do a gym exercise, the second is about administrative formalities, the third comes from a memento, the forth from a cooking recipe. The fifth and the sixth present

instructions about procedures to follow in case of an earthquake, and the two last ones are extracted from a job announcement. Any substitution between these counterparts may give a weird translation and will be inacceptable. Selecting the appropriate Arabic equivalent of the French infinitive entails so treating information concerning text profiles and situational context.

5 Conclusion

In the present article, we examined the transfer of the particularities of the French infinitive into Arabic. We focused on the translation of the infinitive expressing injunctive modality. We showed that a verb in the uninflected infinitive form can be translated by a maṣdar, as an uninflected default counterpart, but it can be also translated by a finite verb. In this case, decisions must be taken concerning the appropriate mood and tense, as well as the right person, number and gender agreement. Resolving anaphoric relations within the text may help but in some contexts, we must refer to the text register or profile to take decision.

The idea behind this research is to demonstrate that grammatically correct translations can be rejected by the use in general. It is therefore necessary to present information concerning usage and situational context in order to get the right translation, this may be ontologically supported. A well formalized machine translation approach must handle this kind of data.

Acknowledgment

I would like to express my gratitude to Igor Mel'čuk for his valuable comments on the preliminary version of this paper.

References

El Kassas Dina. 2007. Vers une classification des équivalents structuraux arabes de l'infinitif français. *Proceedings of the 4th intenational Al Alsun Conference*, April 25-27 2007, Minya: Minya University, Egypt.

Huot, Hélène. 1981. *Construction infinitives du français*. Genève: Droz, Suisse.

Iordanskaja, Lydia, & Mel'čuk, Igor A. 2000. The Notion of Surface-Syntactic Relation Revisited (Valence-Controlled Surface-Syntactic Relations in French). In L.L. Iomdin & L.P. Krysin (eds): *Slovo v tekste i v slovare. Sbornik statej k semidesjatiletiju akademika Ju. D. Apresjana)*, Moskva: Jazyki russkoj kul'tury, 391-433.

Kahane, Sylvain. 2007. A formalism for machine translation in MTT, including syntactic restructurings. *Proceedings of the 3th international conference on Meaning-Text theory*, May 20-24 2007, Klagenfurt: Wiener Slavistisher Almanach, Sonderband 69, Austria.

Mel'čuk, Igor A. 1994. *Cours de morphologie générale*. Montréal: University Press of Montréal, Canada.

Remi-Giraud, S. (dir.) & Roman, Alain. 1988. *L'infinitif : une approche comparative*. Lyon : University Press of Lyon, France.

Vikner, C. 1980. L'infinitif et le syntagme infinitif. *Revue romane*, 15(2): 252-291.

Wagner, R. L. & Pinchon, J. 1991. *Grammaire du français classique et moderne*. Paris : ed. Hachette, France.

Wilmet, Marc. 1997. *Grammaire critique du français*. Louvain-la-Neuve: Duculot, Belgique.

Towards a New Meta-Language for Athabaskan Linguistics: The Case of Morphological Phrasemes

Josh Holden
OLST – Université de Montréal
Département de linguistique et traduction
CP 6128, Succ. Centre-ville
Montréal, Québec, H2C 3J7, Canada
josh.holden@umontreal.ca

Abstract

This paper is concerned with the analysis and citation of morphological phrasemes in the Athabaskan language Dene Sųłiné. The concepts of derivation, morphoids, submorphs and morphological phraseme are reviewed, and a few principles are suggested for distinguishing morphological phrasemes from transparent derivation, followed by corollaries of how these principles can be applied to the presentation of interlinearized examples. This approach is contrasted with the current practice in Athabaskan and Americanist linguistics of not distinguishing between derivation and morphological phrasemes in the morphemic glosses.

1 Introduction

As essential as the concepts of inflection, derivation and etymology are to the analysis of the world's languages, there is no consensus among linguists as to their definitions. While translation practices for wordforms with inflection and derivation markers is fairly straightforward, there is no standard way of interlinearizing those wordforms which, while more analyzable than fully opaque, monomorphemic stems, are too lexicalized to be described as transparent derived stems. The question of how to translate such "morphemes" is particularly relevant when citing interlinearized data from lesser-known languages such as Dene Sųłiné, henceforth Dene, an Athabaskan language from northwestern Canada which is the focus of this paper. Examples (1)-(2), taken from a study of another Athabaskan language[1], are typical of current interlinearization practices in the Athabaskanist literature.

Koyukon (Axelrod 1993:75)

(1) ye–le–∅–tleł
3OB–3.PFV–CL–chop (SEMELFACTIVE)
'he gave it a chop'

(2) le–∅–tseh
3.PFV–CL–cry (SEMELFACTIVE)
'people really mourned his death'

[1] AR="areal" (agreement or argument referring to place or situation), CL=classfier prefix (part of stem), H=high tone, IPFV=imperfective, ITER=iterative, MOM=momentaneous "aspect", NCM=fossilized object noun class agreement, PFV= perfective, REFL=reflexive object agreement, SBEN=self benefactive, VCD=*de*– verb class marker

The semelfactive derivation of a verb indicates an isolated instance of an activity which is usually ongoing. Unlike with (1), however, the meaning of (2) cannot be predicted by combining the signifieds claimed for its component parts. Using the semelfactive form of the verb root meaning 'to cry' should mean 'to cry (once) suddenly', but we see that it has the idiomatic meaning 'to mourn a death'. Although the entire complex is lexicalized, the morphs are still glossed as if they constituted an example of synchronic derivation. What is the best way to present complex wordforms such as (2) so that the difference between morphology and etymology is accessible to researchers not intimately familiar with these languages? Before approaching the presentation of interlinearized examples like (2) it is crucial to review the typology of lexicalized morphological material as understood in Meaning-Text Theory (MTT). If concepts related to inflection, derivation and etymology are vaguely or inconsistently defined, linguists risk speaking at cross purposes when discussing derivational rules or interlinearized data from unfamiliar languages, making them inaccessible to researchers from outside of the specialized field.

When presenting a descriptive analysis of a lesser-studied language, a linguist will divide the signifiers of wordforms into supposedly elementary signs whose signifiers and signifieds could be united according to a descriptive rule to produce the wordform. Elementary linguistic signs are "not representable as [a combination of] other signs united by a meta-operation of linguistic union" (Mel'čuk 1993:64). Elementary signs can typically be classified as affixes or as roots. Following (Mel'čuk 1997:72), affixes mark inflectional and derivational meanings and are distinguished from the root [*racine*], "a stem [which] is a morph or quasi-morph of the language, whose syntactics are similar to the syntactics of most of the morphs or quasi-morphs of L, in that it contains much information about interlexical combinatorics of the wordform it is part of". A large set of unique roots can be combined with at most a few dozen affixes, while the repertoire of up to a few dozen affixes in a language is combined with a very large group of unique stems. The stem [*radical*], simple or derived, "is the part of the wordform **w** that does not include any inflectional affixes which are part of that wordform and which express the grammemes that characterize **w** as a whole" (Mel'čuk 1997:72). Inflectional and derivational affixes are elementary signs with transparent grammatical meanings. Following the criteria in Mel'čuk (1993:311), a derivateme is a standard grammatical meaning which can be productively added to members of a syntactic class, so that speakers can add the derivateme to existing words to coin new lexemes readily interpretable to other speakers. Derivations can be first degree (added to an unanalyzable base, as in *child+ish*, or second degree (childish+ness, institution+al), third degree (institutional+ly), etc. This definition of derivation presupposes that the base is an extant word in the language, so the derived stem must be complex, not elementary. While linguists have often defined derivation in order to distinguish it from inflection (for example, in Bybee 1985 or Anderson 1992), the equally critical distinction between derivation and lexicalized morphological material or visible etymologies has been comparatively neglected.

The above definitions would be sufficient to analyze wordforms if all complex signifiers were linearly divisible into roots, inflectional and derivational affixes, all of which were indisputably elementary signs whose signifieds and signifiers could be transparently combined according to productive morphological rules. Of course, reality is much less neat. For example, should an analyst depict a signifier with an unpredictable meaning such as (2) as complex or elementary? Following Mel'čuk (1993:64), quasi-elementary signs are those whose signifiers are linearly divisible from the rest of the wordform but whose signifieds cannot be isolated from the meaning of the whole word, or alternatively those grammatical signifieds (derivatemes, grammemes) which can be isolated from the meaning of the word although their markers cannot (i.e. a portmanteau in conventional linguistic terminology). This paper is concerned mostly with the first type of quasi-elementary signs, those which are quasi-representable in their signifiers but not in their signifieds. These wordforms are linearly divisible into quasi-elementary signs or quasi-morphs, traces of former morphs whose signifieds cannot be distinguished from the meaning of the whole complex. Following Mel'čuk (1997:246-9) I will use the term *morphoid* to describe quasi-morphs which still bear a plausible semantic link with the meaning of the whole word, and *submorph* to describe those which no longer have so clear a link. In English, the quasi-suffix *–er* in *scanner* and *shaker* is a morphoid because the instrumental signified is still a component of the meaning of the word, regardless of the word's idiosyncratic meaning. A submorph might be the formerly diminutive *–ettes* in the French

lunettes ‘eye glasses’. In no way is *lunettes* related to *lune* ‘moon’ in the modern language. In Dene, many morphoids are visible traces of old derivation markers, old incorporated nouns, and traces of fossilized agreement markers. Complexes of quasi-morphs which are semantically different from the sum of their parts are referred to as morphological phrasemes (Mel’čuk 1997:246-9). One can distinguish inflectional morphological phrasemes, a sort of morphological idiom whereby the language recycles inflection markers to express grammemes other than those typically denoted by those markers, from derivational morphological phrasemes, which are complex signifiers analyzable as an (ex-)derivational morphoid with a supposed lexical base. Such derivational morphoids must have originated as productive derivation markers, but they are by no means chosen by the speaker in the synchronic language, and in some cases would not even be possible to add to their base according to synchronic morphological rules. There must be evidence, however, for the existence at some point of the base to which it was once added. The prefix *re–* in the English verb *rewrite* meaning ‘to change something written in order to improve it’ is such a derivational morphoid, while the same “affix” in *repeat* is not, since *peat* is not a speech verb in English. The visible etymologies of morphoid-base complexes form a continuum between the fully transparent and the opaque. The best analogy in English is the etymology of compounds, which range from the transparent *backstab* (*back+stab*) and *babysit* (*baby+sit*), to the opaque *husband* (Old English *hus* ‘house’ + *bonda* ‘master’) and *garlic* (*gar* ‘spear’ + *leac* ‘root’). The former words were probably coined more recently than the second set, but such distinctions can only be speculative when discussing diachronic derivations in languages without a long written record. When analyzing Dene verbs, one finds a similar range of etymological transparency among complexes of derivational morphoids and former bases for derivation.

Dene verbs are almost exclusively prefixing, so the final element, virtually always the final syllable, is the root. The root is combined with one of four lexicalized prefixes known as “classifiers” in Athabaskanist literature, probably historically valence-changing derivation markers, but which are synchronically part of the stem. The motivation for dividing the classifiers from the root is that the classifiers retain prefix-like phonology such as voicing assimilation of the root’s initial consonant. The root-classifier combination constitutes the simplest type of verb stem, to which an inflectional marker is prefixed (see table 1 below for an outline of the verb’s structure). In addition to these simple, monosyllabic stems, there are discontinuous stems, combining the final radical-classifier pair with one or more elements to the left of the aspect, mood and subject agreement inflectional region. In some cases the origin of these elements is unknown, while in other cases they resemble fossilized prefixes, incorporated adverbs or nouns, or derivational morphoids, giving rise to discontinuous stems. The discontinuous stems can therefore be elementary or quasi-elementary signs.

verb	stem elements	infl	cl	root	gloss
theda		the	Ø	da	‘s/he sits’
heɬxɣ		he	ɬ	ghɣ	‘s/he melts (it)’
k’atheda	k’a	the	Ø	da	‘s/he waits in ambush’
yałtı	ya	Ø	ɬ	tı	‘s/he speaks’

Table 1. Types of Verb Stems in Dene.

The stem elements to the left of the inflection are never chosen by the speaker to coin a new lexeme building on the meaning of the classifier-root pair. The signifier of the verb *k’atheda* ‘s/he waits in ambush’ is partially representable as the formal combination of *k’a* ‘arrowhead’ and *theda* ‘s/he sits’, but the meaning of the element *k’a–* is not analyzable separately from the meaning of the whole verb. It does not denote a way of sitting, since the verb can be used regardless of the physical posture of the subject. The same word is used if the subject is armed with a gun, so ‘arrowhead’ is no longer a semantic

component of the definition. Nor is an element *k'a–* used in any other verbs with an adverbial meaning like 'in ambush'. It is easy to speculate on its diachronic origin as an incorporated noun, but modern Dene no longer has productive noun incorporation. The element *ya–* in *yałtı* is even harder to analyze diachronically. Interlaced with these left-side discontinuous stem elements are genuine transparent derivation markers, including aspect and Aktionsart markers such as the iterative.

Every living language is in a state of constant change, so genuine derivation markers tend to co-exist with formally similar derivational morphoids. So the French derived stems *maison+ette* 'little house' and *fille+ette* 'little girl' are bimorphemic while the morphological phrasemes *lunette* 'eyeglass lens' (*lune+ette *'small moon') and *bicyclette* 'bicycle' are quasi-elementary signs. In Dene morphoids and submorphs are usually former derivational prefixes on a (now quasi-)elementary root.[2] For example, the iterative marker *na–* in Dene can be added to some verbs with the meaning '[V] again'. However a similar morphoid appears in many other verbs which can be considered morphological phrasemes. The meanings of these phrasemes involve repetition in some way, but with other highly specific, unpredictable semantic components. Table 2 provides some examples of both.

Derivation			
base	gloss	base + derivation	gloss
nasther	'I stay'	nanasther	'I stay again'
destth'agh	'I hear (it)'	narestth'a[3]	'I hear (it) again'

Morphological Phraseme			
lexeme 1	gloss	lexeme 2	gloss
noresker	'I'm asking you'	nanoresker	'I'm begging you', *'I'm asking you again'
yarełtı	's/he speaks to self'	k'enayarełtı	's/he repeats self' *'s/he says it again'
yełna	's/he provides for him/her'	nayełna	's/he heals him/her' *'s/he provides for him/her again'

Table 2. Exampes of *na–* the iterative marker and *na–* the morphoid.

The *na–* morphoid in the second set of words obviously resembles the iterative marker, but those lexemes can now be considered unrelated because their meanings are so unpredictable. For example, the verb *k'enáyarełtı* 's/he repeats' is used specifically when someone reiterates his or her idea in a long-winded, verbose way. To indicate saying an utterance twice, a speaker could instead place the adverb *ąłı́ (ıłáh)* '(once) again' before another verb meaning 'say' or 'tell'.

2 Current Presentation of Derivation and Derivational Morphological Phrasemes

Following the tendency of morphologists in other fields, linguists focusing on Athabaskan and many other American languages have given more attention to the relationship between inflection and derivation in their languages of study than to the differences between derivation and derivational morphological

[2] A number of psycholinguistic studies focus on the psychological reality of prefix and quasi-prefix elements. Note that the MTT framework presents principles for building an analytical lexicon that distinguishes between fixed, learned forms and derived words that could potentially be created spontaneously by a speaker who had never heard the word before, and it does not make claims about the psychological reality or processing of these units.

[3] Verb roots are often slightly suppletive (i.e. *–tth'agh/tth'a*) according to the grammemes and derivatemes added.

phrasemes. Athabaskanists tend to use similar terminology for describe complex words like (1) and (2), and tend to present similar morphemic translations for both types. The verbs in (3) and (4) are two more typical examples of the tendency to interlinearize morphological phrasemes as if they were examples of transparent derivation (glosses by the authors)

Slavey (Rice 1989, cited in Rice 2000:144)

(3) dáh–'ede–d–í–d–lu
up–REFL–NCM–ASPECT–CL–handle.rope
'she hanged herself'

Mohawk (Mithun 1984)

(4) t–a–yoti–'nikù:r–v'ne
change–PAST–IT/THEM–mind–fall
'it shocked them'

Example (3) uses a handling verb with a directional prefix *dáh–* 'up' and the reflexive direct object pronominal agreement marker. The noun class marker prefix is a fossilized agreement marker used because the root refers to an object of a particular shape; it is not freely chosen by the speaker. Literally (3) means something like 'she put a rope up for herself' or perhaps 'she tied herself up high', but the actual meaning is 'she killed herself by hanging'. Presumably one could not follow (3) with a second clause like 'but she didn't die'. Example (4) is even more figurative, although the semantic link between the meaning of the word and the morphoids etymologically related to 'mind' and 'fall' is clear. Still, the verb stems in (3) and (4) must mean 'to kill by hanging' and 'to shock emotionally' respectively, with inflectional markers (person agreement and perfective aspect) added.

All of these examples were chosen quite randomly, to illustrate how pervasive a lack of overt distinction between derivation and etymology is, not to single out any particular author or framework. Presenting the data this way does point out some interesting etymologies, and as experts on these languages the authors had no need to indicate what was lexicalized for their own benefit. However, not specifying this knowledge explicitly, and glossing phrasemes as if they were derived stems, risks overgeneralizing the productivity of certain derivational rules, or in the worst cases suggests the existence of derivational rules which are not part of the synchronic grammar. This tendency risks obscuring the morphology to outsiders and ultimately reducing the impact of our studies in the wider linguistics community. The distinction between extant derivation markers and derivational morphoids should therefore be made explicit in the interlinear translation.

Another tendency in Athabaskan and Americanist linguistics is towards breaking polysynthetic wordforms into the smallest identifiable bits, to the point of finding quasi-morphs even where they have apparently not undergone an operation of linguistic union for a very long time. So a noun like *tthot'ɻné* meaning 'English' or 'anglophone' might be interlinearized as *tthe–yeh–hot'ɻné* meaning *stone–house–people:PO.* This is akin to dividing *husband* and *garlic* into *house+bound* and *gar+leek*, compounds recognizable only to students of Old English. It is not enough to discuss such elements as compounds and then note that they are "lexicalized", with no further detail. Presenting synchronic and historical compounds identically risks overstating the productivity or flexibility of compounding in these languages.

A third tendency in Athabaskan and Americanist morphology is to posit "morphemes" based on formal similarities between lexemes. Example (5) from Cook (2004:234) is an example of this.

(5)

ná**dher**	's/he stays'
ts'éni**dher**	'I woke up'
núnɻ**dher**	'it (time) has come'
núniɬ**ther**	'I have grown old'

These verbs share formally similar roots. The author discusses these verbs as "lexically derived" from a common abstract root, whose "literal meaning", it is claimed, "refers to passing time". While at one point the author mentions that words are related "at least etymologically", they are nonetheless discussed as an example of synchronic derivation. While it is fascinating to speculate about such etymological relationships, defining "morphemes" on the basis of formal similarity can lead to odd and unprincipled divisions. For example, Singh and Neuvel (2003) cite need to avoid giving morphemic status to elements such as an *eau* 'water' "morpheme" common to *bat**eau*** 'boat' and *rad**eau*** 'raft', despite their semantic commonality. Similarly situated are the famous phonosthemes of Firth (1930), such as the *gl–* sequence common to ***gl**itter*, ***gl**eam*, ***gl**isten*, etc. Unfortunately when lesser-known languages are being analyzed by speakers of European languages, this trap may be easier to fall into. This has the additional disadvantage of leading the analyst to posit unusual grammatical rules in these languages that allow for the derivation of a small set of such forms. In Dene, the existence of perhaps a few dozen words similar to *k'atheda* 's/he waits in ambush' has resulted in Athabaskan grammars presenting this language as having noun incorporation of instrumentals and direct object arguments, albeit admittedly different from classic incorporation in languages such as Inuktitut and Chukchi. While noun incorporation need hardly be completely productive to be considered part of the synchronic grammar, even the most "lexical" sorts of incorporation (Mithun's Type I incorporates) involve the coining of new verb-incorporate combinations as an available synchronic process, at least until the languages are critically endangered (Mithun 1984). In Dene, on the other hand, verbs with incorporated nouns form a small repertoire of learned forms. So in Athabaskan studies conflating derivation with derivational morphological phrasemes has resulted in Dene grammar as being characterized as having incorporation, while the existence in French of similar words like *maintenir* 'maintain, support' (from *main* 'hand'+ *tenir* 'hold') and *colporter* 'carry on neck' (*col* 'neck' + *porter* 'carry') has not led linguists to posit noun incorporation as part of French grammar, because the distinction between studying derivation and studying etymology in European languages has remained relatively stark.

3 Improving Consistency Between Interlinearization and the Grammar

It is particularly vital to distinguish synchronic derivations from morphological phrasemes in the presentation of interlinearized wordforms and texts. Current interlinearization practices in Americanist and Athabaskan studies reflect the aforementioned tendency to analyze and gloss any sequence that could possibly be considered linearly divisible as a sequence of morphs. The analyst decides how small the divisions should be based on his or her knowledge of the lexicon and etymology or for the convenience of the unrelated purpose at hand. Below are three examples of how the morphemic glosses of two Dene phrasemes might be presented if they appeared in various major grammars or dictionaries.

	nayełna 's/he heals him/her'	*nanoresker* 'I beg you'
	na–ye–ghe–ł–na	na–ne–ho–de–s–ł–ker
(6)	ITER–3SG.OB–3SG.SUBJ–CL–live	ITER–2SG.OB–AR–VCD–1SG.SUBJ.IPFV–CL–root
(7)	heal–3SG.OB–3SG.SUBJ–CL–heal	ITER–2SG.OB–ask –1SG.SUBJ.IPFV–CL–ask
(8)	3SG.OB/3SG.SUBJ.IPFV:heal	2SG.OB/1SG.SUBJ.IPFV:beg

The differences are not due to conflicting morphological analyses so much as to vague definitions. No difference is made between learned forms with origins as derived stems and derivatemes chosen by the speaker. While certainly the clearest way to present examples of transparent sychronic derivation and inflection, this approach crashes hard when applied to morphological phrasemes. Within a grammar or dictionary, various discontinuous stems might not be broken down and glossed to a similar degree, depending on the analyst's knowledge of and interest in etymologies. This is confusing for students of morphology and etymology as well as for linguists not intimately familiar with Athabaskan morphology.

To ensure a more transparent presentation of interlinearized examples, it would be helpful to consistently apply the following principles whatever the specific decisions of the analyst may be.

A. Derivations must be distinguished from morphological phrasemes.

B. Morphoids, submorphs and elementary signs should all be distinguished from each other.

The first principle is more important than the second, because treating elementary and quasi-elementary signs identically mischaracterizes the grammatical system of the modern language. The second principle is highly valuable, because in erring on the opposite side and presenting long verb stems with an array of morphoids as unanalyzable elementary signs we risk losing valuable information about diachronic change and lexicalization. Below are three corollaries of the above principles, showing how they could be applied to real interlinear translations.

1. In an interlinearization of a morphological phraseme, morphoids and their (former) morphemic glosses are italicized, while derivation markers and their glosses are not. The root or base is also italicized, so everything italicized is synchronically lexical rather than grammatical. All elements not italicized are part of the synchronic grammar. Compare (9a) with (9b).

(9) a. narestth'a
na–*de*–s–*∅*–*tth'a*
ITER–*hear*–1.IPFV–*CL*–*hear*
'I hear again'

b. nanoresker
na–ne–*hode*–s–*ɬ*–*ker*
ITER–2SG.OB–*ask*–1.IPFV–*CL*–*ask*
'I beg you'

Derivational morphoids are part of the stem, and are only glossed as former grammatical morphemes if the analyst can posit a diachronic evolution of that derived stem into a morphological phraseme. This restriction also helps distinguish derivational processes that exist in modern Dene, such as Aktionsart derivation, from processes that are theorized to have existed in the past but are longer active.

2. While all noninflectional and nonderivational material is italicized in the second line, only morphoids are glossed. Submorphs are not glossed at all (comparing *na–* in 10a and 10b).

(10) a. nayełna
na–ye–ghe–*ɬ*–*na*
ITER–3OB–3SUBJ.IPFV–*CL*–*live*
'he is healing him

b. nabadhi
naba–*∅*–*∅*–*dhi*
nosy–3.IPFV–*CL*–*nosy*
'she is nosy'

Naturally it is up to the linguist to decide which elements are morphoids and which are submorphs, i.e. what constitutes a plausible semantic link. Even *nabadhi* 's/he is nosy' in (10b) and the verb *badhi* 's/he wants to ingest (it)' share the semantic component of desire, so one could argue that *na–* is a morphoid rather than a submorph. An analyst could go further and speculate that *badhi* originated as *ba* 'for' + *dhi* 'think', although *–dhi* alone is not a stem denoting 'think' in modern Dene. Assuming that all these elements are considered submorphs rather than morphoids, they should be mentioned outside of the four-line interlinearization format, for example in a footnote. A verb stem with an "incorporated" noun like *k'atheda* in Table 1 would be glossed as quasi-elementary (with the etymology in a footnote), as in (11).

(11) k'atheda
k'a–the–*∅*–*da*
wait.ambush–3.IPFV–*CL*–*wait.ambush* (SG)
's/he waits in ambush'

3. Morphoids can be divided off until the analyst reaches an indivisible base. Any linear division of a morphological phraseme implies a diachronic meta-operation of linguistic union.

To treat elements as derivational morphoids, there must be an extant (ex-)base in the language to which the derivation marker could have been affixed. This prevents a wordform like (12a) from being entirely divided into morphoids (apart from its inflection markers), none of which could have been a base for derivation, as in the alternative gloss in (12b).

(12) a. nanoresker
na–ne–*hode*–s–*ɬ*–*ker*
ITER–2SG.OB–*ask*–1SG.IPFV–*CL*–*ask*
'I beg you'

b. nanoresker
na–ne–*ho*–*de*–s–*ɬ*–*ker*
ITER–2SG.OB–*AR*–*VCD*–*CL*–*stem*
'I beg you'

The portion *–hode–* in (12a) could in principle be further divided as in (12b), but the remainder *reɬker/ɬker* is not a meaningful sign, so no meta-operation of linguistic union could apply. In English, this corollary allows for a division like *betray+al* but not *be+tray+al*. At least a plausible story of diachronic linguistic union must be possible. This keeps the practice of glossing morphoids consistent with the MTT definition of derivation in Section 1. Moreover, glossing the portion *hode..ɬker* as a stem meaning 'ask' in (12a) constitutes a claim that this sequence is fact an existing Dene verb stem (true).

Applying this principle consistently to Dene verbs produces some rather unconventional interlinearizations, given the unusual interlaced ordering of discontinuous stem elements, inflectional and derivational prefixes in Athabaskan verb morphology. Whatever is glossed as the stem must be constant across the inflectional paradigm. The lexical meaning of the verb is associated with the whole stem including the morphoids.

(13) natthírest'a
natthíre–s–*∅*–*t'a*
get.up–1.IPFV–*CL*–*get.up* (IPFV)
'I get up (in the morning)'

(14) tunéhdą
tune–H–*∅*–*dą*
drown–3.PFV–*CL*–*drown*
'he drowned'

(15) hasonéɬtą
há–se–*hone*–H–*ɬ*–*tą*
teach–1OB–*teach*–3.PFV–*CL*–*teach* (PFV)
'they taught me'

The stem of (13), could in principle be given the etymology *na–tthí–de...∅–t'a* or *ITER–head–NCM...CL–move.round*, literally something like 'to move (as a round object) one's head again', just as etymology of the stem meaning 'to drown' in (14) could be *tu–ne...∅–dą* or *water–MOM...CL–drink*, literally 'to drink water to an endpoint'. It is harder to posit an etymology for (15), but one could speculate about the origin of several of the submorphs. According to the conventions proposed in this paper, all of these speculative etymologies should be relegated to footnotes or somewhere outside of the interlinear gloss, so as to clearly differentiate them from the morphology. The present author has already been applied to this style of interlinear translation to long Dene texts, making the morphology easily accessible to non-Athabaskanists and even without overburdening the text with footnotes.

Conclusion

Interlinear translation forces the analyst and the reader to confront the interaction between the language's morphology and its lexicon. Linguists working outside of MTT may disagree on the precise definitions of concepts such as the morpheme, derivation, phraseology, etc., and their relationship with the lexicon, and each framework has its own traditional terminology. However, all linguists should agree that a specific analysis is better than avoidance of such thorny questions. The MTT concepts of morphological phraseme, quasi-elementary and elementary signs, morphoids and submorphs provide an extremely valuable framework for presenting interlinearized translations of lexicalized wordforms with visible etymologies without undermining the grammatical description. The translation practices suggested above could be extremely helpful in explaining the extremely numerous morphological phrasemes in understudied languages of theoretical interest such as Dene to the larger linguistics community. The MTT framework therefore has the potential to enhance the contribution of Athabaskan linguistics to the broader fields of American linguistics and to general morphology and lexicography.

Acknowledgements

I would like to express my heartfelt thanks to the Dene speakers who assisted me with this and other research, especially Jesse Sylvestre, Greg Noltcho, Jerry Noltcho, Gilbert Benjamin, and Elmer Campbell, all of Buffalo River Dene Nation, as well as the staff of the Dene Dánarílnai Foundation and the other residents of Dillon. The suggestions and criticism of Igor Mel'čuk, David Beck and Jasmina Milićević were also extremely helpful.

References

Anderson, Stephen. 1992. *A-morphous morphology*. Cambridge: Cambridge University Press.

Axelrod, Melissa. 1993. *The Semantics of Time. Aspectual Categorization in Koyukon Athabaskan*, Lincoln, Nebraska: University of Nebraska Press.

Bybee, J. 1985. *Morphology: A Study of the Relation Between Meaning and Form.* Amsterdam/Philadelphia: John Benjamins.

Cook, E-D. *A Grammar of Dene Sųłıné (Chipewyan)*. Winnipeg: Athabaskan and Iroquoian Linguistics.

Firth, J. R. 1930. *Speech*. London: Benn's Sixpenny Library.

Mel'čuk, I. A. 1993. *Cours de Morphologie Générale* (volume 1). Montreal: Presses de l'Université de Montréal.

Mel'čuk, I. A. 1997. *Cours de Morphologie Générale* (volume 5). Montreal: Presses de l'Université de Montréal.

Mithun, M. 1984. The evolution of noun incorporation. Language, 60 (4), 847-895.

Rice, K. 2000. M*orpheme Order and Semantic Scope. Word Formation in the Athapaskan Verb*, Cambridge: Cambridge University Press.

Singh, R and S. Neuvel. Quelles unités? Les unités morphologiques. In B. Fradin *et al.* (eds.), *Silexicales 3*. Villeneuve d'Ascq - Université de Lille 3.

Foresight or Hindsight: the Mystery of Russian *spoxvatit'sja*

Boris Iomdin
V.V.Vinogradov Russian Language Institute, Russian Academy of Sciences
Volkhonka 18/2, 119019, Moscow, Russia
iomdin@ruslang.ru

Abstract

The paper deals with the Russian verb *spoxvatit'sja*, hard for translation. The verb is a unique example of three different mental activities (memory, perception, and comprehension) fused in a single Russian word. It has a peculiar set of seemingly quite different and even opposite meanings which turn to be organised in a logical polysemy structure. It also has a variety of interesting syntactic features, partly shared by other Russian verbs denoting mental acts.

1 How many senses?

Consider the following examples taken from parallel corpora:

(1a) *He **wondered** now why he had not asked whether this train went on through to Montreal or some Canadian point.* (Theodore Dreiser. Sister Carrie)

(1b) *А Герствуд вдруг **спохватился**, что даже не спросил, идет ли этот поезд прямо до Монреаля или до какого-нибудь другого пункта в Канаде.* (Translated by M. Volosov)

(2a) *Tom **found himself** writing "BECKY" in the sand with his big toe; he scratched it out, and was angry with himself for his weakness.* (Mark Twain. The Adventures of Tom Sawyer)

(2b) *Том **спохватился**, что пишет на песке "Бекки" большим пальцем ноги; он стер написанное и рассердился на себя за такую слабость*. (Translated by N. Daruzes)

(3a) *I think he hardly knew what he was saying, for when I asked him what business he was in he answered, "That's my affair," before he **realized** that it wasn't the appropriate reply.* (F. Scott Fitzgerald. The Great Gatsby)

(3b) *Должно быть, он думал в это время о чем-то другом, потому что, когда я спросил его, чем, собственно, он занимается, он ответил "Это мое дело", а только потом **спохватился**, что ответ был не очень вежливый.* (Translated by E. Kalashnikova).

(4a) *Your grammar is—" She had intended saying "awful," but she **amended it to** "is not particularly good".* (Jack London. Martin Eden)

(4b) *Вы говорите...– она собиралась сказать "ужасно", но **спохватилась** и докончила: – не очень грамотно.* (Translated by R. Oblonskaya)

In all four occasions, the translators used the Russian word *спохватиться* for what appear to be quite different meanings. When analysing more contexts with this word, however, we can indeed see that it has all of these meanings (at least).

In most traditional Russian explanatory dictionaries, though, no clear distinctions are made. Consider e.g. the definition of *спохватиться* in (Evgenyeva 1981-1984): "*Colloquial*. Remember smth. suddenly, discover a committed oversight or blunder"; similar definitions are given in other acknowledged dictionaries of Russian. Bilingual dictionaries treat the verb in a similar way; consider e.g. "*(coll.)* to remember suddenly, think suddenly" (Oxford Russian Dictionary); "*coll.* sich besinnen (*вспомнить*), plötzlich bemerken (*заметить*)" (Rymashevskaya 1999); "se ressaisir, se reprendre, s'aviser; se rappeler brusquement de qch (*вспомнить*); s'apercevoir de qch (*заметить*)" (Shcherba et al. 2002); "*coll.* caer en la cuenta, darse cuenta (repentinamente); ocurrirse de repente (*вспомнить*); advertir de repente (*заметить*)" (Turover & Nogueira 2001). As we can see, the bilingual dictionaries in fact distinguish two different meanings of the verb, labeling them with the Russian synonyms *вспомнить* 'to remember' and *заметить* 'to notice'. These two meanings, although not explicitly divided, can also be found in the definition of the explanatory dictionary given above.

Indeed, if we consider (1) and (2), we can say that they exemplify these two meanings: 'remember' in (1) and 'notice' in (2). But does remembering or noticing actually take place in (3) or (4)? A person could not possibly *remember* that his own reply was inappropriate unless he had forgotten it before, which is clearly not the case in (3). Hardly could he *notice* it, either, since noticing is paying attention to something as it is happening and not afterwards. Moreover, in (4) the character surely couldn't *remember* or *notice* what she was saying, since she hadn't said it yet!

So, for examples like (3), one could propose the meaning 'understand'. Here is another example of such usage:

(5a) *Moreover, as if **perceiving** at last that if he should give undiluted conscientious advice to Pip, he would be leaving him too wide a margin to jump in for the future, Stubb suddenly dropped all advice.* (Herman Melville. Moby Dick)

(5b) *Вот почему, словно **спохватившись**, что, давая Пипу советы на совесть, во всей их неразбавленной сложности, он только подсказывает ему оправдание для всех его будущих прыжков в воду, Стабб вдруг прервал свои поучения*. (Translated by I. Bernstein)

Indeed, Stubb could neither *remember* that he should not give advice (since he clearly never thought of it before), nor *notice* it (since it was a new idea that just came up in his own mind).

For examples like (4), 'check oneself' would be a more appropriate English equivalent. Cf. (6) and (7):

(6a) *'And perhaps you were never even introduced to a lobster–' (Alice began to say 'I once tasted–' but **checked herself** hastily, and said 'No, never')…* (Lewis Carroll, Alice's Adventures in Wonderland)

(6b) *– И, должно быть, никогда не видала живого омара... – Зато я его пробова... – начала Алиса, но **спохватилась** и покачала головой. – Нет, не видала.* (Translated by N. Demurova)

(7a) *A fire of mingled love and the passion of gambling came into Daylight's eyes. Involuntarily his hand started for his pocket for the coin. Then it **stopped**, and the light in his eyes was troubled.* (Jack London. The Burning Daylight)

(7b) *Глаза Харниша загорелись любовью и игорным азартом. Рука невольно потянулась к карману за монетой. Но он тотчас **спохватился**, и взгляд его затуманился.* (Translated by V. Toper)

2 *Вовремя спохватиться* vs. *поздно спохватиться*

In (4), (6), and (7), the verb *спохватиться* means 'half doing something undesirable and changing one's mind in time not to do it', which is also neither remembering nor noticing. When using the verb this way, speakers of Russian often add the adverbial modifier *вовремя* 'in time'; cf. (8)

(8a) *Having supposed that there was sense where there is no sense, and a laudable ambition where there is not a laudable ambition, I am well out of my mistake, and no harm is done.* (Charles Dickens. The Tale of two Cities)

(8b) *Я имел неосторожность предположить здравый смысл и похвальное честолюбие там, где их нет и в помине, но* ***вовремя спохватился*** *и счастливо избежал ошибки, так что все обошлось как нельзя лучше.* (Translated by S. Bobrov & M. Bogoslovskaya)

Note that in (8a), no fragment corresponds directly to *вовремя спохватиться* in (8b): the idea of timely canceling something planned lies in the combination of the gerund (*having supposed*) and the phrase *be well out of* in the main clause. This helps us understand the semantic structure of the Russian verb in this sense. It has a presupposition of intending to do something and then understanding that this was undesirable or inappropriate, and an assertion of not doing it. Hence, its presupposition is what the assertion is in (3) or (5). To show this, we can add the negative particle *не* and see what happens:

(9a) *Несколько дней спустя Энгельсон, нисколько не думая о том, что он говорит и при ком, сказал: – На такую мерзость способен только немец. Гауг обиделся. Энгельсон уверял его, что он* ***не спохватился****, что у него сорвалась эта глупость нечаянно с языка.* (А. И. Герцен, Былое и думы)

(9b) *A few days later Engelson who did not at all think of what he was saying or who was listening to him, said: Only a German could do such a nasty thing. Haug was offended. Engelson tried to assure him that he just* ***did not check himself****, that he let the nonsense slip.* (Alexander Herzen. My Past and thoughts)

(10a) *Художник засмеялся и даже* ***не спохватился****, что, может, грешно смеяться-то. Не увидел, не заметил, не обратил внимания, какой стоял Смородин…* (В. Шукшин, Пьедестал)

(10b) *The painter laughed and* ***didn't*** *even* ***realize*** *that it probably was a sin to laugh. He didn't see it, didn't notice, didn't pay attention to how Smorodin looked–* (Vassily Shukshin. The pedestal)

In (9a), *не спохватился* means that he did not refrain from saying what he thought inappropriate although he did realize that it was inappropriate, while in (10a), it means that he did not realize at all that laughing was inappropriate. Thus, in (9) the verb is used like in (4), (6), and (7) and means 'check oneself', while in (10) it is used like in (3) or (5) and means 'understand'.

Levontina (2004) notices that the adverb *вовремя* ['in time'] very often modifies verbs that describe situations when one is planning to do something or starts doing it, but later decides not to do it: *спохватиться, одуматься* ['think better of it'], *сдержаться* ['restrain oneself'], *прикусить язык* ['bite one's tongue']. Note, however, that one can also say *поздно* ['tardily, too late'] *спохватиться*, but not **поздно сдержаться <удержаться* ['refrain'], *прикусить язык>*. This is due to the fact that *сдержаться*, *удержаться*, *прикусить язык* mean 'resist the temptation to do something inappropriate', and a combination of this meaning with 'too late' leads to a semantic paradox. At the same time, *поздно спохватиться* is perfectly standard, since here *спохватиться* means 'realize'. Cf. 11:

(11a) *That was the other thing. No cigarettes. Oh well.* ***Too late now****. I'll do that tomorrow.* (Helen Fielding. Bridget Jones's Diary)

(11b) *Вот еще чего нельзя. Никаких сигарет... Ах, черт, ладно.* ***Поздно спохватилась****. С завтрашнего дня брошу.* (Translated by G. Bagdasaryan)

We observe here an interesting assymetry. Normally, *вовремя* is opposed to *рано* 'too early' or *поздно* 'too late', cf. (Levontina 2004: 112). One can say *рано приехать* 'come too early', *вовремя приехать* 'come in time' or *поздно приехать* 'come too late'. With verbs like *сдержаться* or *удержаться*, only *вовремя* can be used, since their semantic representations include the meaning 'in time': they contain the idea of changing one's mind after one thinks of something, but before he actually does it. And with *спохватиться*, one can say both *вовремя спохватиться* and *поздно спохватиться*, but the verb would mean different things here (unlike *приехать* in the examples before, which means 'come' in all of them). Moreover, in highly colloquial speech one can even say *рано спохватиться*, which would mean the same thing as *вовремя спохватиться*; cf. 12:

(12a) *Доктор наш нами (нашим прогрессом в лечении) очень доволен. Говорит, что молодцы, что* ***рано спохватились*** (From a medical web forum).

(12b) The doctor was much satisfied with us (with the progress of the treatment). He tells us that we did a great thing that we **reacted early**.

Here, *рано спохватились* does not mean 'realized earlier than needed', but 'identified the disease and visited the doctor before it was too late'. Interestingly, the peculiar phrase can also be used in its direct meaning: in (13) the Russian translator of a short novel by John Collier uses it to convey the irony of the speaker, who blames his friend for noticing fire and putting it out before his old house burns down and he gets the insurance:

(13a) *"Angry!" Said Mark laughing. "We are though, for putting it out. Why didn't you let the confounded place burn down?* (John Collier, Great possibilities)

(13b) *– Нет, мы очень на вас сердиты – за то, что вы* ***рано спохватились****. Дали бы этому постылому домине спокойненько догореть.* (Translated by M. Makarov)

3 Four senses: the invariant and the differences

We have seen that beside the meanings of 'remember' and 'notice' given in the dictionaries, the verb *спохватиться* has at least two other meanings: 'understand' and 'check oneself'. But can we say that all dictionaries are unaware of these two meanings or that these meanings have nothing to do with those described by the definitions? Not really. In fact, all the uses of the verb that we have analysed have much in common. They share the following semantic components:

(a) 'undesirable situation for which the subject is to blame': this is conveyed in the dictionaries by words like *упущение* 'oversight', *промах* 'blunder', etc.
(b) 'idea of this situation coming into the subject's mind': the dictionaries put it as *вспомнить* 'remember' or *обнаружить* 'notice'.
(c) 'fact that the situation has to be dealt with before it's too late'.

The differences lie in how these components are organized in the semantic representation of each lexeme of *спохватиться*.

Below, tentative definitions are proposed for each of the lexemes:

Sense 1: ‘A person suddenly remembered something and realized that an undesirable situation, into which s/he should have intervened earlier, has been taking place’.

Sense 2: ‘A person suddenly discovered something and realized that an undesirable situation, into which s/he should have intervened earlier, has been taking place’.

Sense 3: ‘A person suddenly understood something and realized that an undesirable situation, into which s/he should have intervened earlier, has been taking place’.

Sense 4: ‘Having realized that an undesirable situation is going to happen if the person does what s/he has been intending or starting to do, do something else instead’.

We can see that, basically, there are two points in which these meanings differ: (*i*) why the idea of the undesirable situation comes into the subject’s mind (by recollection, discovery or comprehension of something) and (*ii*) when it comes to him (before the situation has developed or only afterwards). This brings us to two interesting observations.

4 Semantic observation: *спохватиться* as a unique amalgam of mental activities

The first opposition provides us with a very interesting example of three different mental activities (memory, perception, and comprehension) fused in a single Russian word. The idea that, in the linguistic picture of the world, these subsystems of our mind work in close cooperation is not new (for Russian, see e.g. Apresjan 2001, 2008; Arutjunova 2000; Bragina 2007; Dmitrovskaja 1988, 1991; Iomdin 2006, Iomdin to appear; Kubryakova 2004; for other languages, many interesting observation and further bibliography in Language of Memory 2007), but it has hardly ever been mentioned that one and the same word can express the meanings of ‘remember’, ‘discover’, and ‘understand’. Moreover, in some examples one could hardly distinguish which of the meanings should be attributed to the verb usage. Cf.

(14a) *– Заболталась, а вы есть хотите, –* ***спохватилась*** *повариха.* (А. Рыбаков, Дети Арбата)

(14b) *I keep talking, and you are hungry, – the cook* ***interrupted herself****.* (Anatoly Rybakov. Children of the Arbat)

(15a) *Они успели проехать несколько кварталов, когда она вдруг* ***спохватилась****: – А куда, собственно говоря, ты меня везешь?* (А. Маринина, Мужские игры)

(15b) *It was not until they had passed several blocks that she suddenly* ***exclaimed****: Hey, where are you driving me to?* (Alexandra Marinina. Male games)

Did the cook in (14) remember that people she spoke to were hungry, or did she understand it from their looks, or did she just glance at the clock? Did the woman in (15) remember a plan, or did she understand that something unforeseen was happening, or did she notice that the she was being driven along unknown streets? All this is unclear from the verb *спохватиться* and could only be understood in a broader context.

5 Syntactic observation: government patterns

The differences in the semantic representations correspond to the differences in syntactic features, namely the valency structure and government patterns. *Спохватиться* can have two different valency structures. In senses 1, 2, and 3, it has two semantic actants: A1 ‘Subject’ and A2 ‘Situation’. Syntactically, the Subject is expressed in trivial ways, while the Situation valency can be instantiated by a sentence, either

introduced by the conjunction *что* (consider examples 1, 3, 5, 9, 10) or without it, as a direct speech (either before the verb, as in 14, or after it, as in 15). In sense 4, the verb has another semantic actant: A3 'Action of the Subject' (e.g. what the Subject did instead of the action he refrained from). Surprisingly enough, this valency cannot be instantiated in ordinary ways, such as a noun phrase in the Instrumental case (with or without prepositions): one cannot say **спохватиться ответом <молчанием, кивком головы>* [literally 'by a reply, by silence, with a nod'] nor **спохватиться в приветствии <в извинении>* [literally 'in welcoming words, in an excuse'], as one says *отреагировать молчанием <кивком головы, письмом>* ['react by silence <with a nod, by a letter'], *выступить с приветствием <с извинением>* ['make a welcome <apologetic> speech']. Neither can one implement this actant by a clause: **спохватиться тем, что P*. The only two ways of instantiating this valency are the following:

(a) introducing it with *и* 'and': cf. *спохватилась и докончила* in (4), *спохватилась и покачала головой* in (6), or *спохватился и счастливо избежал ошибки* in (8).
(b) putting the verb *спохватиться* into a gerund clause and expressing the valency A3 by the main clause, as in (16):

(16a) *Лиманский машинально потянулся к уху, намереваясь поправить несуществующую дужку, но,* ***спохватившись****, сделал вид, что почесывает затылок.* (Е. Прошкин, Механика вечности)

(16b) *Limansky absent-mindedly reached out for his ear to adjust the non-existent earpiece, but* ***catching himself*** *pretended that he was scratching his head.* (Evgeny Proshkin, Mechanics of eternity)

These two exotic ways of instantiating a valency were first described by Igor Boguslavsky for verbs *изловчиться* 'contrive, manage somehow', *поднатужиться* and *поднапрячься* 'to make big efforts and succeed'. Cf.

(17a) *Шариков в это время* ***изловчился*** *и проглотил водку*; *– У самих револьверы найдутся… – пробормотал Полиграф, но очень вяло и вдруг,* ***изловчившись****, брызнул в дверь* (М. Булгаков, Собачье сердце)

(17b) *Sharikov* ***seized this moment*** *to gulp down his vodka*; *You're not the only one with a revolver… muttered Poligraph quietly. Suddenly he* ***dodged*** *and spurted for the door*. (Mikhail Bulgakov, The Heart of a Dog, translated by Michael Glenny)

Boguslavsky also mentions that this usage is possible in other languages, including English and Swedish:

(18) *We were too hungry even to try and think of anything except food.* (George Orwell, Down and out in Paris and London)

(19a) *Var så god och svara på frågen*

(19b) *Please answer the question*, literally 'Be so good and answer the question'
(Boguslavsky 1996: 28–32).

Note that the verbs described by Boguslavsky denote physical actions, while *спохватиться* is a mental act. It turns out that there are at least two other mental verbs for which this government pattern is the only possible one: these are *забыться* 'forget oneself' and *зазеваться* 'stand gaping, let one's thoughts wander and miss something'. Cf.

(20a) ***Забывшись****, я по трактирной привычке принялся вытирать тарелку салфеткой.* (М. Шишкин, Всех ожидает одна ночь)

(20b) ***Forgetting myself****, I started to wipe the plate with a napkin as if I was in a local.* (Mikhail Shishkin, One night awaits everyone)

(21a) *Зархи не любил свежую зелень в супе, и ЛЮ каждый раз боялась* ***забыться*** *и бросить ему щепотку укропа.* (В. Катанян, Лиля Брик)

(21b) *Zarkhi did not like fresh herbs in his soup, and LY always feared that she would* ***forget herself*** *and throw in a pinch of dill in his plate.* (Vassily Katanyan, Lilya Brik)

(22a) *Маша,* ***зазевавшись****, однажды угодила ногой в один из капканов.* (А. Мусатов, Стожары)

(22b) *Walking and* ***gaping****, Masha once got her foot right into the trap.* (Alexey Musatov, Stozhary)

(23a) *На севере Юты мы* ***зазевались и*** *проскочили нужный нам поворот.* (В. Песков, Б. Стрельников, Земля за океаном)

(23b) *Traveling somewhere in Northern Utah, we were* ***nodding and*** *missed the right turn.* (Vassily Peskov & Boris Strelnikov, The land overseas)

In these two verbs, the valency in question does not express a desirable situation (as in *спохватиться* or in Boguslavsky's examples), but an undesirable one. In (Iomdin, to appear) we propose the following definitions for these verbs (A2 is the valency in question):

Забыться 'The person A1, ceasing to be aware of conditions that have to be observed in the current situation, violated them by doing A2'

Зазеваться 'Having focused on something unrelated to the current situation, the person A1 failed to notice something important happening, which caused an undesirable situation A2'

Moreover, for the Russian verb *вспомнить* 'to remember', the semantic valency of Content can not only be instantiated in standard ways (most of which are impossible for *спохватиться*[1]), cf. *вспомнить дату <число, лицо>* 'remember the date <number, face>', *вспомнить о невыполненном обещании <про полученное письмо>* 'remember about the unfulfilled promise <about the received letter>', *он вспомнил, что надо было позвонить отцу <где лежат деньги>* 'he remembered that he had to ring his father <where the money was>', but also using the construction with the conjunction *и* 'and'; cf.

(24a) *Я даже ждал, что Алексей однажды позвонит. Просто* ***вспомнит и позвонит.*** (В. Маканин, Андеграунд, или герой нашего времени);

(24b) *I even expected that Alexey would once call me. That he just* ***remembers to call****.* Vladimir Makanin. Underground, or a hero of our time)

(25a) *Потом, стоя уже в прихожей, чтобы закрыть за ним дверь, она* ***вспомнила и попросила*** *его ввинтить лампочку в ванной.* (А. Кабаков, Девушка с книгой, юноша с глобусом, звезды, колосья и флаги).

[1] In older dictionaries one can find examples of the transitive usage of the verb, cf. ***Спохватилась*** *мачеха пасынка, когда уже лед прошел* ['The stepmother did not **miss** her stepson (who presumably drowned in winter) until the ice drifted'], but this usage is now obsolete, and one should say *хватиться* 'miss, notice the absence' instead.

(25b) *Then, already standing in the entryway, she* ***remembered she meant to ask*** *him to fix a lamp in the bathroom.* (Alexander Kabakov. A girl with a book, a boy with a globe, stars, ears of grain and flags. Translated by Natasha Perova)

This is possible if one speaks about remembering a necessity (to ring, as in (24), or to ask to change the bulb, as in (25)). Here, *вспомнить и сделать А2* [literally 'remember and do A2'] means 'having remembered that A2 had to be done, do A2'. This construction is symmetrical to its opposite, a far more standard construction *забыть сделать А2* 'having forgotten that A2 had to be done, not to do A2', cf.

(26a) *Доктор, эти господа, вероятно, второпях,* ***забыли положить*** *пулю в мой пистолет.* (М. Ю. Лермонтов, Герой нашего времени)

(26b) *Doctor, these gentlemen* ***have forgotten****, in their hurry, no doubt,* ***to put*** *a bullet in my pistol*. (Mikhail Lermontov, A hero of our time. Translated by J. H. Wisdom & Marr Murray)

Anna Zalizniak argues that the English sentence *I remembered to ring Bill* cannot be literally translated, since the verb *вспомнить* is impossible here: one cannot say ***Я вспомнил позвонить Биллу* but can only say *Я не забыл позвонить Биллу* 'I did not forget to ring Bill'. Zalizniak explains this difference of English and Russian usage by the fact that "*Vspomnit'* implies that the object was forgotten for a certain period of time (a semantic component which is absent from *I remembered to ring Bill*)" (Zalizniak 2007: 99). However, it seems that this component is not necessarily contained in the meaning of *вспомнить*. It can be absent if one speaks about imagining an object or person which has never disappeared from one's memory. Cf.

(27a) *He got back in fancy to the old Hurstwood, who was neither married nor fixed in a solid position for life.* (Theodore Dreiser. Sister Carrie)

(27b) *Он вспомнил свою молодость, когда он еще не был женат и не имел еще прочного места в жизни.* (Translated by M. Volosov)

Besides, it seems that this semantic component can be present in the English construction as well, cf. (28):

(28) *Suddenly I remembered to ask Stephen how he'd found me* (Gabrielle Roy, Enchantment and sorrow. Autobiography of Gabrielle Roy).

So, in this situation one does use *вспомнить*, if the Content valency is instantiated using the conjucntion *и*: *Я вспомнил и позвонил Биллу* [literally 'I remembered and rang Bill']. Cf. also (29), when this construction is used twice, both for *спохватиться* and for *вспомнить*:

(29a) *Ты вдруг* ***спохватишься и вспомнишь и*** *успеешь меня поздравить с днем Химика* (А. Битов, Письмо).

(29b) *You would suddenly* ***think and remember and*** *congratulate me in time on Chemist Day.* (Andrey Bitov, A letter)

6 Conclusions

We have seen that the seemingly very different meanings of *спохватиться* have in fact much in common. Taking into account its semantic and syntactic properties (the semantic emphasis on realizing that an undesirable situation could or did happen rather than on how exactly it was realized, different

presuppositions and assertions and government patterns), one could sum up the usage of the verb into two different meanings (or lexemes, in the sense of Moscow Semantic School), but not those two provided by the dictionaries. We propose the following definitions for these lexemes:

Спохватиться 1: 'A person A1 suddenly realized that an undesirable situation A2, into which s/he should have intervened earlier, has been taking place'.

Спохватиться 2: 'A person A1, having realized that an undesirable situation A2 is going to happen if A1 does what s/he has been intending or starting to do, does A3 instead'.

Such a set of meanings, as it seems, is quite interesting and not common for other European languages. Still we would refrain from drawing any conclusions for the uniqueness of the Russian linguistic picture. The idea of realizing something too late, conveyed by the first lexeme, can also be expressed by the idiom *задним умом крепок* ['wise behindhand', literally 'strong with the hind mind'], and is also typical for Polish (*mądry Polak po szkodzie*), Czech (*pozdě bycha honí*, *mysli mu to až třeti den*), Bulgarian (*късно му идва умът*), English (*wise after the event, it is easier to have hindsight than foresight, deathbed repentance*), German (*die besten Gedanken kommen hinterher*, *hinterher sind alle Dummen schlau*), French (*l'esprit de l'escalier*), Spanish (*es estratega a posteriori*), Italian (*dietrologìa, del senno di poi ne son piene le fosse, il giudizio vien tre giorni dopo la morte, dopo il fatto ognuno è savio*), and probably many other languages; see also (Zelenin 2002). The second lexeme can also be quite easily translated into most languages. It is the combination of the two in one and the same word that makes it worth analyzing.[2]

Acknowledgements

This work was done with a support by the Program of Fundamental Research "Genesis and interaction of social, cultural, and linguistic communities" of the History and Philology Department of Russian Academy of Sciences and by the Russian President grant for the Support of Leading Scientific Schools No. НШ-3205.2008.6. Examples from the National Corpus of Russian Language (www.ruscorpora.ru) were used.

References

Amberber, Mengistu (ed.) 2007. *The language of memory in a crosslinguistic perspective*. John Benjamins Publishers.

Apresjan, Juri D. 2001. *Системообразующие смыслы 'знать' и 'считать'*. In: *Russkij jazyk v nauchnom osveschenii*. No.1:5–26. [*The system forming meanings 'know' and 'believe'*. In Russian.]

Apresjan, Juri D. 2008. *Systematic Lexicography*. Oxford University Press.

Arutjunova, Nina D. 2000. *Знать себя и знать другого (По текстам Достоевского)*. In: *Slovo v tekste i v slovare*: Festschrift for the 70th anniversary of Juri Apresjan. Moscow. Shkola "Jazyki russkoj kul'tury". [*Knowing oneself and knowing another person. Based on Dostoevsky texts*. In Russian.]

Boguslavsky, Igor M. 1996. *Сфера действия лексических единиц*. Moscow. Shkola "Jazyki russkoj kul'tury". [*The scope of lexical units*. In Russian.]

Bragina, Natalya G. 2007. *Память в языке и культуре*. Moscow. Jazyki slavjanskix kul'tur. [*Memory in language and culture*. In Russian.]

Dictionary of Russian Folk Dialects. 1999. Ed. 33. S. Petersburg. Nauka.

[2] Amusingly enough, the peculiar combination of meanings corresponds to the combination of prefixes (*с-* and *по-*), both with a perfective meaning, which is very uncommon for Russian. For the first meaning, it would be more natural to use the prefix *про-* which can mean 'do something unimportant and miss the main thing', cf. *прозевать, проморгать, прохлопать, прошляпить, проворонить*. And, indeed, one finds the word *прохватиться* in Russian dialects; cf. ***Прохватился*** *брат, да поздно* ['The brother **thought better of it** but it was too late'] (Dictionary of Russian Folk Dialects: 24).

Dmitrovskaya, Maria A. 1988. *Знание и достоверность*. In: *Logicheskij analiz jazyka. Pragmatika i problemy intensional'nosti*. Moscow. Institute of Oriental Studies, Russian Academy of Sciences. 166–188. [*Knowledge and veracity.* In Russian.]

Dmitrovskaya, Maria A. 1991. *Философия памяти*. In: *Logicheskij analiz jazyka. Kul'turnye koncepty*. Moscow: 83–92. [*Philosophy of memory.* In Russian.]

Evgenyeva, Anastasia P. 1981–1984. *Dictionary of Contemporary Russian Language*. Moscow. Russkij jazyk.

Iomdin, Boris L. 2006. *Языковая модель понимания*. In: *Jazykovaja kartina mira i sistemnaja leksikografija.* Ed. by Juri D. Apresjan. Moscow. Jazyki slavjanskix kul'tur. 515–612. [*Linguistic model of comprehension*. In Russian.]

Iomdin, Boris L. To appear. *Ментальная лексика: память и ее функционирование*. In: *Aktivnyj slovar' russkogo jazyka. Prospekt i slovnik*. Ed. by Juri D. Apresjan. [*Mental lexicon: memory and how it functions*. In Russian.]

Kubrjakova, Elena S. 2004. *Язык и знание*. Moscow. Jazyki slavjanskix kul'tur. [*Language and knowledge*. In Russian.]

Levontina, Irina B. 2004. *Вовремя, своевременно*. In: *Novyj ob"jasnitel'nyj slovar' sinonimov russkogo jazyka*. Ed. by Juri D. Apresjan. 2nd edition. Moscow. Jazyki slavjanskix kul'tur. Wien. Wiener Slawistischer Almanach. 112–115. [*In time, timely* [a series of synonyms]. In Russian.]

Oxford Russian Dictionary. 2007. Oxford University Press.

Rymashevskaya, Emilia L. 1999. *Modernes Deutsch-Russisches Wörterbuch*. Moscow. Russkij jazyk.

Shcherba, Lev V., Margarita I. Matusevitch, & Sofya A. Nikitina. 2002. *Grand dictionnaire russe-français*. Moscow. Russkij jazyk.

Turover G. Ya. and J. Nogueira. 2001. *Gran diccionario ruso-español.* Ed. by G. Ya. Turover. Moscow. Russkij jazyk.

Zalizniak, Anna A. *The conceptualization of remembering and forgetting in Russian*. In: *The language of memory in a crosslinguistic perspective*. Ed. by M. Amberber. John Benjamins Publishers. 97–18.

Zelenin Alexander. 2002. *Русские задним умом крепки. А другие народы? Фразеологизм как отражение национального характера*. In: *Russkij jazyk za rubezhom*. No.2:45–50. [*The Russians are wise after the event. What about other nations? The idiom as a reflection of national character*. In Russian.]

Linguistic Well-formedness of Semantic Structures

Lidija Iordanskaja
OLST – Université de Montréal
C.P. 6128 Centre-ville
Montréal, H3C 3J7 Canada
lidija.iordanskaja@umontreal.ca

Igor Mel'čuk
OLST – Université de Montréal
C.P. 6128 Centre-ville
Montréal, H3C 3J7 Canada
igor.melcuk@umontreal.ca

Abstract

The paper formulates the problem of language-specific constraints on linguistically well-formed Semantic Structures: how can we draw a borderline between semantically well-formed and ill-formed sentences, and how can we tell lexical incongruence from semantic ill-formedness? The constraints that determine semantic ill-formedness in a given language are based on a simple criterion: If the meaning of a linguistically anomalous sentence can be expressed by a linguistically correct paraphrase, then the anomaly is not semantic; otherwise, it is semantic. Several relevant examples (in English, French and Russian) are analyzed; a sketch of a typology of semantic constraints within the Meaning-Text framework is proposed. We also define the distinction between extralinguistic and linguistic well-formedness of SemSs and propose a calculus of possible cases.

1 The Problem Stated

This paper deals, within the Meaning-Text framework, with the transition

{ Semantic Structures [= SemSs] } ⇔ { Deep-Syntactic Structures [= DSyntSs] },

examined from one particular angle:

> How to specify the linguistic well-formedness of the input set, that is, the set of SemSs for a particular language **L**?

Logically, there are two major types of well-formedness of the representational structures adopted in this framework: purely formal well-formedness (satisfying the general, language-independent conditions imposed on Meaning-Text semantic networks and trees) and linguistic well-formedness (satisfying language-specific conditions). This paper considers only **linguistic well-formedness of SemSs**.

Specifying the output set—that is, DSyntSs—has been tackled before. In this domain, Apresjan's work (1978, 1989) advanced some deep ideas, introducing two types of syntagmatic constraints on well-formedness of DSyntSs—constraints based on semantic restrictions. Apresjan showed that the absurdity of the starting SemS leads to linguistic anomaly only if the expression of a contradictory combination of meanings includes 1) contradicting grammemes/grammaticized lexemes or 2) contradicting lexemic presuppositions/modal frames. Here we will try to develop his ideas as applied to the level of SemSs—namely, language-specific syntagmatic constraints on combinations of semantemes, that is, on Sem-configurations. (A semanteme is, roughly speaking, the signified of a full lexical unit [= LU] of given language **L**.)

Speakers judge SemSs mostly proceeding from actual sentences. If a sentence is anomalous—odd or outright incorrect, we want to see at what level of representation this anomaly can be detected. According to Apresjan, the distinction must be drawn between extralinguistic "distortions" (he considered contradictions and tautologies) and strictly linguistic anomalies. We concentrate on a further distinction, namely

the distinction between two types of linguistic anomaly: semantic *vs.* lexical-syntactic. This distinction rests on the following simple criterion (slightly reformulated from Apresjan, 1978 [1995: 609]):

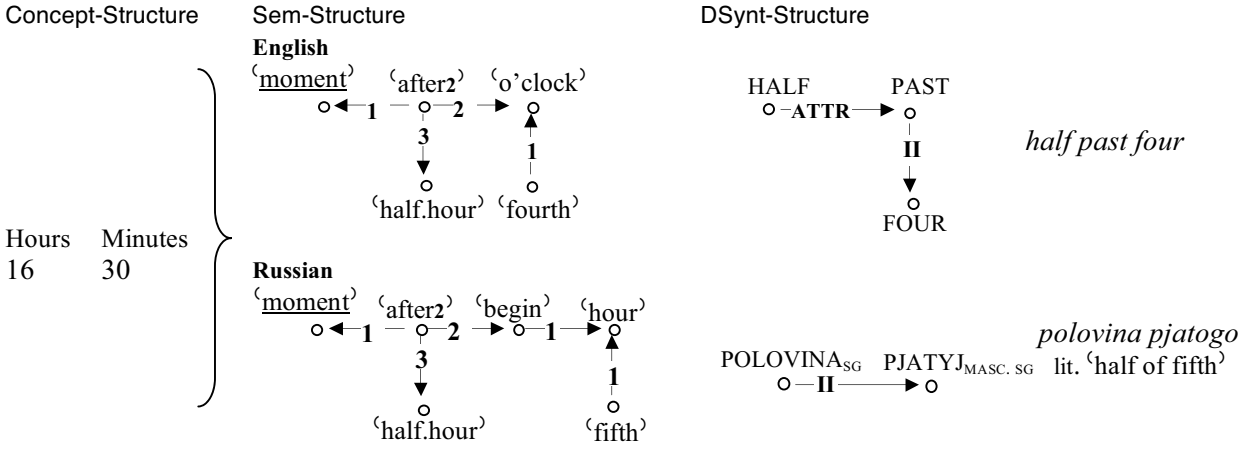
> If the meaning ‘σ’ of a linguistically anomalous sentence can be expressed by a linguistically correct paraphrase without modifying ‘σ’, then the anomaly is not semantic; otherwise, that is, if, in order to obtain a correct paraphrase, ‘σ’ has to be replaced with another, although (quasi-)equivalent meaning ‘σ′’, the anomaly is semantic.

Two different meanings, and, consequently, two different SemSs are (quasi-)equivalent if they have the same denotation.

Let us emphasize, lest a misunderstanding arises, that by “a linguistically correct expression” we mean an expression that is normally acceptable in everyday non-technical speech.

To make our basic idea clear, we will begin with an example. Namely, we will present the main structures at the three levels of representation for two semantically equivalent expressions of English and Russian: Concep(tual)S(tructure), SemS and DSyntS. The expressions chosen for illustration are equivalent phrases *half past four* and *polovina pjatogo*, both naming a particular moment during the day: “16: 30.”

In order to save space, we give semantemes of Russian in English (even if this contradicts the nature of semantemes, which are language-specific; we do the same with French semantemes later on); the distinguishing lexicographic numbers are taken from *Longman Dictionary of Contemporary English Online*.



The semanteme ‘o’clock’ means ‘one of 24 moments that divide a day in equal parts’, and the semanteme ‘hour’ is ‘duration between two consecutive o’clocks’; the ordinal numeral corresponds—in this case—to the order of ‘o’clock’/‘hour’ after ‘noon’, i.e., after the moment when the Sun is at the highest point (rather than to the quantity of ‘o’clocks’ and ‘hours’); finally, the underscoring in SemSs specifies the communicatively dominant node.

Figure 1. SemSs and DSyntSs for an English and a Russian expression naming a moment.

As seen from Fig. 1, the same informational content, in our case, the identification of a particular moment, can be expressed in two different languages by two non-isomorphic SemSs—that is, by two different meanings (which are of course equivalent). The moment in question is named in English by giving the time interval that has passed from the 4-th o’clock till this moment (‘half.hour after the 4-th o’clock’), while in Russian this is done by indicating a moment in the middle of the full hour following the 4-th o’clock (‘half of the fifth hour’). Thus, English and Russian use different meanings in order to name the same moment: in English, the point of reference is ‘the moment “the fourth o’clock”’, while in Russian, it is ‘the moment of the beginning of the fifth hour’. (Cf. B. Russell’s famous example of the two different meanings corresponding to the same referent: ‘Sir Walter Scott’ and ‘the author of *Waverley*’.)

Different SemSs corresponding to the same extralinguistic content are quite a common phenomenon in natural language; such correspondences are found between different languages as well as inside the same language. What makes such SemSs interesting is the fact that some of them can be good in one language/in one context, but bad in a different language/in a different context. Thus, the English SemS of *half past*

four is not allowed in Russian (**polovina posle četyrëx*), and vice versa: **half of* [*the*] *fifth* is impossible in English. One can speak of well- or ill-formed SemSs of a **given language**, aiming at their **linguistic** well-/ill-formedness—namely, at Sem-configurations that are correct/incorrect in a given language. As a result, the following problem arises:

> What are the constraints that ensure, in a given language, the specification of linguistically well-formed SemSs and only such SemSs?

Section **2** proposes several such constraints, in particular—for the case presented in Fig. 1.

2 Constraints on Configurations of Semantemes

We decided to explore language-specific constraints on well-formed SemSs for the three following reasons:

1) Formally, it is necessary to specify as accurately as possible the set of objects with which our model is supposed to deal as its input.

2) Practically, it seems more economical, under synthesis, to filter out bad sentences as soon as possible—at the semantic source. This is processing economy.

3) Linguistically, it allows for significant economy in the lexicon. Restrictions on cooccurrence attached to a semanteme are valid, as a rule, for all LUs in whose meaning this semanteme is the dominant component; and the number of such LUs can reach tens of thousands. Such is the case, for instance, with the semanteme ‹cause**1**› and all causative verbs (see **2.5**). Otherwise, we would have to repeat all these restrictions with each individual LU.

A bad SemS in a language **L** is bad, of course, only because **L** lacks necessary lexical-syntactic means for its implementation. Therefore, it is logically possible to do without constraints on Sem-configurations —by means of constraints on Lexicalization. From the linguistic point of view, however, this is an unacceptable solution, because it misses an essential generalization: "The meaning of such a structure is **never** lexicalizable in **L**."

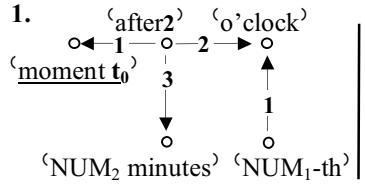
Now we will examine several cases illustrating linguistic well-/ill-formedness of Sem-configurations.

2.1 Naming a Moment

The meaning corresponding to "16: 30" cannot be expressed in English as **half of* [*the*] *fifth*, which is a literal rendering of the Russian pattern. It is impossible to save this expression by superficial corrections: **half of the fifth hour* or **thirty (minutes) of the fifth (hour)* are equally unacceptable, although syntactically these expressions are impeccable and 100% understandable. What is at stake here is not a bad realization of a well-formed SemS, but an ill-formed SemS. To specify some well-formed English and Russian SemSs in the case where the moment named does not correspond to a round hour (= to an ‹o'clock›), we need three semantemic patterns:

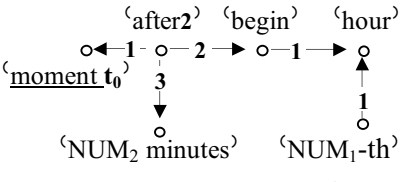
English **Russian**

1.



1) NUM_1 = last o'clock before t_0
2) $NUM_2 < 30$

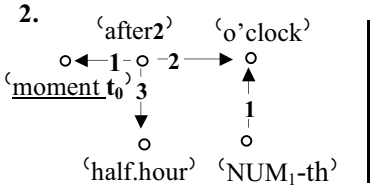
(*twenty past four*)



1) NUM_1 = first o'clock after t_0
2) $NUM_2 < 30$

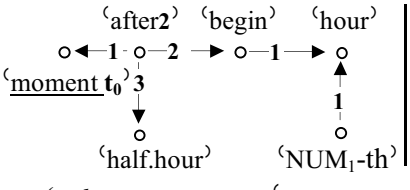
(*dvadcat' minut pjatogo* lit. ‹twenty minutes of fifth [hour]›)

2.



1) NUM_1 = last o'clock before t_0
2) $NUM_2 = 30$

(*half past four*)



1) NUM_1 = first o'clock after t_0
2) $NUM_2 = 30$

(*polovina pjatogo* lit. ‹half of fifth [hour]›)

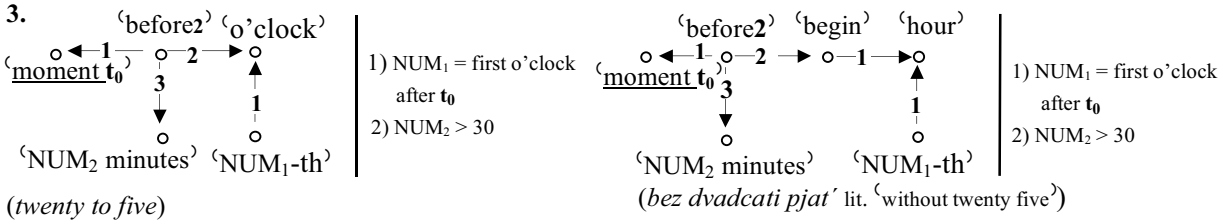


Figure 2. English and Russian Semantemic Patterns for Denoting a Time Moment.

The corresponding constraint must be attached to the semanteme ‹moment› in both languages:

> A SemS containing a Sem-configuration of the form ‹moment←1–after2/before2–3→minutes› is allowed to be used in order to name a moment only if this configuration is part of one of the three above patterns for a given language.

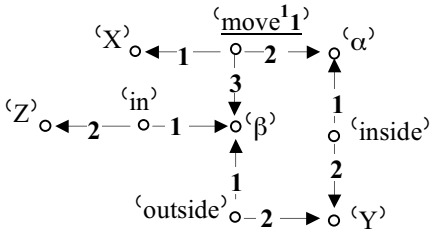
2.2 Moving Out from Somewhere to Somewhere

Consider the following familiar example (see, e.g., Apresjan, 2000: xii-xiii), the French sentence (1):

(1) *Le serpent est sorti de son trou* ‹The snake came out of its hole›.

Our object of interest here is the verb SORTIR ≈ ‹X comes out of Y to Z› = ‹X moves^{1}**1** from point α, which is inside Y, to point β, which is outside Y—in Z›; this definition is represented by the SemS in (2):

(2)



As is well known, the SemS (2) cannot be implemented in non-technical, everyday Russian; the sentence (3) is unacceptable:

(3) a. **Zmeja peremestilas´ iz nory naružu* lit. ‹[The] snake moved^{1}**1** out.from [its] hole outside›.

The linguistic reason for this is also known: Russian requires that, for any act of moving^{1}**1**, the way of locomotion be specified—that is, ‹by walking›, ‹by flying›, ‹by crawling›, ‹by jumping›, etc.[1] Moreover, if the movement is characterized by its beginning point or its end point, this additional obligatory Sem-component must be normally expressed inside a single lexeme—usually, a derived verb of the type VY+JTI ‹come out by walking›, VY+LETET´ ‹come out by flying›, VY+POLZTI ‹come out by crawling›. Thus, (3b) is also unacceptable as a description of the fact described in (1), the correct way being (3c):

b. **Zmeja peremestilas´ iz nory naružu polzja* lit. ‹The snake moved^{1}**1** out of its hole outside by crawling›.
c. *Zmeja vypolzla iz nory* ‹The snake crawled out of its hole›.

(The expression of the correct Sem-configuration ‹move^{1}**1** by crawling› by a single lexeme in Russian is ensured under the lexicalization of the SemS.)

Therefore, Russian needs the following semantic constraint:

[1] The verb PEREMESTIT´SJA ‹change place› as in *Sekretariat peremestilsja v sosednee zdanie* ‹Secretaries moved to the next building› is a different lexeme. PEREMESTIT´SJA in (3) can be normally used for generic statements in scientific-type texts: for instance, *Ulitka peremeščaetsja so skorost´ju 0,08 m/min* ‹A snail moves with the speed of 0.08 m/min›.

> A SemS containing the semanteme ‹move^{1}**1**› is allowed in non-technical everyday speech only if ‹move^{1}**1**› is part of the Sem-pattern ‹move^{1}**1←1**–manner–**2→Ψ**›.

In contrast, French avoids expressing the manner of locomotion together with the indication of movement's orientation in general—if there is no special communicative need to insist on it. One can say

(4) *Le serpent est sorti de son trou en rampant* lit. ‹The snake came out of its hole by crawling›,

but this puts a communicative emphasis on ‹crawling›. The corresponding constraint for French is:

> A SemS containing a Sem-configuration of the form ‹move^{1}**1**–**2→**α› or ‹move^{1}**1**–**3→**β› is allowed in non-technical everyday speech if 1) ‹move^{1}**1**› is not part of the pattern ‹move^{1}**1←1**–manner–**2→Ψ**› or 2) if the manner of locomotion is informationally especially valuable for the Speaker.

Both constraints are to be associated to the corresponding semantemes: Fr. ‹se déplacer = move^{1}**1**› and Rus. ‹peremeščat´sja = move^{1}**1**›. In case we are translating from French into Russian, we have to select one of more specific synonyms of ‹peremeščat´sja = move^{1}**1**›, whose meanings include the semanteme of manner of locomotion: for instance, ‹vypolzti = move^{1}**1** out by crawling›, ‹vybežat´ = move^{1}**1** out by running›, etc. As a result, the French sentence (1) is translated into Russian as (3c).[2]

The phenomenon at stake here is what is called conflation (Talmy, 2003: 25 and *passim*): different Sem-configurations are covered by different LUs (within one language or across languages).

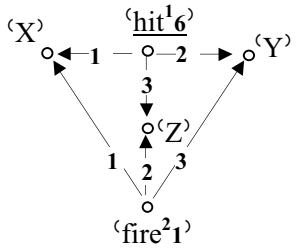
2.3 Shooting Somebody

The English sentence (5) cannot be easily translated into Russian:

(5) *Mary shot John*.

This happens because Russian has no verb exactly corresponding to SHOOT^{1}**1** ‹X shoots Y with Z›, which has the lexicographic definition ‹X hits^{1}**6** Y with projectile Z which X fires^{2}**1** at Y›; the decomposition of the semanteme ‹shoot^{1}**1**› can be represented by the SemS in (6):

(6)



HIT^{1}**6** ≈ ‹reach object Y with projectile Z›; to simplify the presentation, we omitted a SemA of HIT^{1}**6** (and, correspondingly, of SHOOT^{1}**1**): hit where? (*John was shot/hit* ***in the leg***).
FIRE^{2}**1** ≈ ‹send projectile Z at Y›.

The closest Russian equivalent of this SHOOT, the verb STRELJAT´, means ‹X fires **at** Y› (= shoots^{1}**2** at Y): it does not imply hitting the target, which constitutes the semantically dominant component of ‹shoot^{1}**1**›, cf. (6). SemS (6) can be literally expressed in Russian as (7):

(7) *X popadaet v Y-a Z-om, vystreliv v Y-a Z-om* ‹X hits^{1}**6** Y with Z having fired^{2}**1** Z at Y›.

Nevertheless, none of sentences in (8) is acceptable as a translation of (5):

(8) a. #*Mèri popala v Džona, vystreliv ⟨streljaja⟩ v nego* ‹Mary hit^{1}**6** John firing^{2}**1** ⟨having fired^{2}**1**⟩ at him›.
b. #*Mèri vystrelila ⟨streljala⟩ v Džona i popala (v nego)* ‹Mary fired^{2}**1** at John and hit^{1}**6** him›.

All these sentences are syntactically and morphologically correct; no lexical cooccurrence is violated, either. The only conclusion is, then, that the SemS of (5) is ill-formed in Russian.

[2] A similar difference between motion and position verbs in French and German is well illustrated in (Malblanc, 1961: 66*ff*).

However, it is impossible to ban the SemS of the form (6) from Russian altogether, because in particular contexts it can be implemented:

(9) a. *Mèri popala v zajca, streljaja s kolena* ‹Mary hit^{1}**6** the hare from the kneeling position›.
b. *Mèri vystrelila v zajca i, kak ni stranno, popala*
‹Mary fired^{2}**1** on the hare and, strange though it may seem, hit^{1}**6** him›.

SemS (6) can be allowed in Russian only if the speaker is especially interested in the "intermediate" result of firing—namely, hitting the target. In a neutral situation of firing at Y, the goal is to harm Y; reaching this goal implies hitting, but not the other way around (for instance, Y can be protected by good body armor). Therefore, sentences in (8) are informationally deficient with respect to the meaning of (5): they tell us more about Mary's actions and "successes," while (5) is about what has happened to John. This can be captured by the following constraint, associated with the Russian semanteme ‹streljat′ ≈ fire^{2}**1**›:

> A SemS containing a Sem-configuration of the form (6) is allowed only if the meaning ‹hit^{1}**6**› is informationally especially valuable for the Speaker.

English avoids this problem because in (5) the semanteme ‹hit^{1}**6**› is not expressed by a separate LU.

Since the SemS in (6) cannot be used in Russian to describe the situation present in (5), semantic paraphrasing is triggered—namely, looking for quasi-synonyms of the dominant node of (6), the semanteme ‹hit^{1}**6**›. Its Russian equivalent, ‹popadat′›, has more specific quasi-synonyms ‹ranit′ ≈ injure› and ‹ubit′ ≈ kill›, one of which must replace Sem-configuration (6)—if Y is a living being. The result is (10a) or (10b):

(10) a. *Mèri ranila Džona* ‹Mary injured John›.
b. *Mèri ubila Džona* ‹Mary killed John›.

These are, of course, very approximate translations of (5): they lose information about firing, but, at the same time, they add information about the final result of the firing. This means that the translator needs knowledge about the real extralinguistic situation (he has to know what actually happened to John).

2.4 Be Localized Somewhere

Sentence (11a) cannot be rendered in Russian literally—as (11b); the correct form is (11c):

(11) a. *Mary stiffened up beside John ⟨in her chair⟩.*
b. **Mèri vsja naprjaglas′ rjadom s Džonom ⟨v svoem kresle⟩.*
c. *Mèri, kotoraja byla rjadom s Džonom ⟨sidevšaja v kresle⟩, vsja naprjaglas′*
‹Mary, who was beside John ⟨sitting in her chair⟩, stiffened up›.

The problem with (11b) is caused by the fact that the Russian semanteme ‹be.localized› (the dominant node in the decomposition of any locative preposition)—in contrast to the English one—cannot accept as its SemA **1** the designation of a particular class of states such as some uncontrollable human physiological reactions (stiffen up, go limp, shake). The corresponding restriction must be attached to the Russian semanteme ‹be.localized›. Because of this constraint, the Sem-configuration ‹stiffen.up←**1**–beside [Y]› is ill-formed in Russian; the rules of Sem-paraphrasing will modify it to produce a different, but equivalent SemS, in which ‹beside [Y]› will bear on ‹Mary›, so that the SemS for (11c) is as follows:

Rus. ‹John←**2**–beside–**1**→Mary←**1**–stiffened.up›.

2.5 Causing Something

The English semanteme ‹cause**1**› (= ‹be the cause of›) takes as its SemA **1** the designation of an object impacting something; this is true of any transitive verb whose meaning's dominant node is ‹cause**1**›:

(12) a. *The hammer falling from the balcony broke the window*.
b. *John's bullet killed the wolf instantly*.

But sentences in (13) are bad:[3]

(13) a. *John struck the window with a hammer, and* #*the hammer broke it*.
b. #*John's rifle killed the wolf instantly* [for the situation in which John fired at the wolf].

ʿCause**1**ʾ cannot take the name of an object as its SemA **1**, if, in the situation described, this object is an instrument being used by somebody (Kahane & Mel'čuk, 2006); this restriction must accompany the semanteme ʿcause**1**ʾ in English (and also in French and Russian).

In the same vein, (14a) cannot be literally rendered in Russian as (14b); the correct way is (14c):

(14) a. *An ambush killed four soldiers.*
b. Rus. **Zasada ubila četyrëx soldat.*
c. Rus. *V zasade pogiblo četyre soldata* ʿFour soldiers died in an ambushʾ.

The Russian semanteme ʿcause**1**ʾ has an additional restriction on its SemA **1**: at least, in the situation of killing, the cause X (of the death) must be in physical contact with Y that undergoes the causation; ʿcause**1**ʾ is the dominant node of the corresponding semanteme ʿkill**2**ʾ. Therefore, the SemS

ʿambush←**1**–kill**2**ʾ

is ill-formed in Russian, and must be replaced by an equivalent one:

ʿambush←**2**–use–**1**→α←**1**–kill**1**–**2**→soldiersʾ (≈ ʿα kills**1** soldiers by using an ambushʾ),

which eventually gives (14c).

3 Semantic *vs.* Lexical-Syntactic Linguistic Ill-formedness

The borderline between ill-formed semantemic combinations (in a SemS) and ill-formed lexical combinations (in a DSyntS) is not always obvious. Let us illustrate the respective cases.

A sentence synthesized from a well-formed SemS may be ill-formed at the DSynt-level for one of two reasons: either an incorrect choice of LU for a correctly used semanteme, but without breaking the rules of mutual cooccurrence of LUs; or an incorrect choice of LU that leads to impossible or undesirable lexical cooccurrence.

- **Bad lexical choice: Rus. IZ-ZA ≈ ʿbecause ofʾ**

Russian sentences (15a) and (15b) clearly contrast, (15b) being pragmatically bizarre:

(15) a. *Ivan pogib iz-za svoej rassejannosti* ≈ ʿIvan died because of his absentmindednessʾ.
b. #*Ivan spassja iz-za svoej rassejannosti* ʿIvan was saved because of his absentmindednessʾ.

The explanation is clear: (15b) is bad because *spassja* ʿwas savedʾ is, generally speaking, a desirable event, while the preposition IZ-ZA ʿ≈ because ofʾ implies the absence of desirability of X for X's corresponding SemA or for the Speaker (Iordanskaja & Mel'čuk, 1996: 169–172):

ʿX iz-za Y-aʾ = ʿX because of Y, it being not true that X is desirable for $SemA_1$(ʿXʾ) or for the Speakerʾ.

[3] They are bad in a non-contrastive context only; cf. *John failed to break the window until he struck it with a hammer; the hammer broke it instantly* or *My pistol had no effect on the wolf, but John's rifle killed the animal*.

The meaning of (15b) is then as follows: ‘Ivan was saved because of his absentmindedness, it being not true that being saved is desirable for Ivan or for the Speaker’. In spite of its pragmatic strangeness, this meaning is well-formed: the maxim “to be saved is good” is not a semantic fact of a particular language, but rather a piece of encyclopedic information. One can even imagine a situation in which being saved is not considered as something desirable (for instance, Ivan wanted to die, but failed). The problem with sentence (15b) lies in a lexical choice: if the Speaker does not want to negate the desirability of Ivan’s being saved, he has to express the semanteme ‘because of’ by the preposition BLAGODARJA ‘thanks to’:

‘X blagodarja Y-u’ = ‘X because of Y, X being desirable for $SemA_1$(‘X’) or for the Speaker’.

The problem of the correct selection of LUs in accordance with the starting SemS and the knowledge and intentions of the Speaker is orthogonal to the problem of correct combination of semantemes in accordance with their properties. The anomaly in question cannot be blocked by a syntagmatic filter of the DSynt-level: no lexical cooccurrence restrictions are violated in (15b). The result of a bad lexical choice of this type is always a relatively incorrect sentence (incorrect with respect to a given meaning). Its pragmatic unacceptability is due to the fact that the wrongly selected LU may add some unwarranted information, which renders the sentence odd.

• Bad lexical combination: Rus. NEMNOGO ≈ ‘a little’ and Rus. RUSYJ ‘light-brown [hair]’

We will present two examples of bad lexical combinations banned by syntagmatic constraints (more precisely, semantic conditions) on lexical cooccurrence.

— As shown in (Iordanskaja & Mel’čuk, 2004), the Russian adverb NEMNOGO ≈ ‘a little’ does not combine with a “threshold” lexeme L—a lexeme whose meaning includes the component ‘sufficiently bigger/smaller than the norm’. The semanteme ‘sufficiently’ determines the threshold after which L can apply. Thus, one can be *nemnogo grusten* ‘a little sad’, but not **nemnogo vesel* ‘a little cheerful’: one is called *grusten* as soon as he is even slightly below the norm of neutral emotional state, but to be called *vesel*, one has to be **sufficiently** over the norm—one has to cross a particular threshold. Yet the combination of the semanteme ‘nemnogo = a little’ with a threshold semanteme is quite possible, although in such a case, the lexeme NEMNOGO itself cannot be used:

for RANEN ‘injured’	: ‘nemnogo ranen’	⇔	***legko*** *ranen* ‘lightly injured’	⟨**nemnogo ranen*⟩
for SOGLASEN ‘agree’	: ‘nemnogo soglasen’	⇔	***otčasti*** *soglasen* ‘agree partially’	⟨**nemnogo soglasen*⟩
for ŽËLTYJ ‘yellow’	: ‘nemnogo žëltyj’	⇔	*želt*+***ovat***-*yj* ‘yellowish’	⟨**nemnogo žëltyj*⟩

The observed ungrammaticalities are due to the violation of lexical restricted cooccurrence, not ill-formed SemSs; thus, ‘nemnogo vesel’ can be correctly expressed as *vesel, no ne očen´* ‘cheerful, but not much’ or *ne sliškom vesel* ‘not too cheerful’. Of course, the phrase **nemnogo vesel* also illustrates bad lexical selection; however, this anomaly, in contrast to the preceding case, is blocked syntagmatically. The semanteme ‘a little’ corresponds to the lexical function [= LF] `AntiMagn`, and the choice of the proper element of an LF is done at the surface-syntactic level.

— Many languages use special adjectives to denote the color of human hair and human eyes. Thus, Russian has ***rusye*** *volosy* ‘light-brown hair’ and ***karie*** *glaza* ‘brown eyes’, and these adjectives cannot be applied to anything else; their definitions are as follows: ‘rusye X’ = ‘light-brown hair X’ and ‘karie X’ = ‘brown eyes X’. As a consequence, the expressions **karie volosy* ‘brown hair’ and **rusye glaza* ‘light-brown eyes’ are ungrammatical, but this ungrammaticality corresponds to violation of lexical cooccurrence. SemSs ‘light-brown hair’ and ‘brown eyes’ are well-formed; however, for hair and eyes the brown color cannot be expressed by the adjective (SVETLO-)KORIČNEVYJ ‘(light-)brown’: **svetlo-koričnevye volosy* ‘light-brown hair’ and **koričnevye glaza* ‘brown eyes’ are bad phrases in Russian because of the restricted combinability of particular LUs.

Bad lexical combinations produce absolutely incorrect sentences.

Thus, semantic constraints on Sem-configurations must be consistently distinguished from semantic conditions on lexical combinations (Iordanskaja & Mel’čuk, 2004: 122*ff*). Two LUs units can be non-combinable only because of their meanings; but this does not entail that their meanings are non-com-

binable. Very often two semantemes are perfectly combinable, but one of them must be lexically expressed as a function of the lexical implementation of the other; any violation results in lexical-syntactic ill-formedness. This is the case of lexical functions.

4 Typology of Semantic Well-formedness Constraints

The data analyzed allows us to sketch a preliminary typology of linguistic constraints that come into play at the semantic level and specify linguistically well-formed SemSs.

A linguistic constraint attached to a particular semanteme ‘σ’ of **L** can be of two types—as a function of the constraint's target: this is either the Sem-Governor of ‘σ’/of one of SemAs of ‘σ’ or the Sem-Dependent of ‘σ’; in other words, either the semanteme ‘σ_G’ of which ‘σ’ (or a SemA of ‘σ’) is a SemA, or the semanteme ‘σ_D’ which is a SemA of ‘σ’. These two types can be dubbed "up-oriented" and "down-oriented."

1. Up-oriented semantemic constraints. Two subtypes are distinguished: semantic clichés and lexical conflations.
   1.1. Sem-clichés: Sem-configurations to be expressed by lexical-syntactic schemata (see **2.1**: "naming a moment").
   1.2. Lexical conflations: Sem-configurations to be expressed by single LUs or to be avoided, since there is no good lexical expression, that is, "positive" or "negative" conflations—some semantemes in a language are inseparable, while in another language they can be incompatible **(**see **2.2** and **2.3**: "moving out from somewhere" and "shooting somebody").

2. Down-oriented semantemic constraints. The data examined so far show also two subtypes: the constraint bears either only on one SemA of ‘σ’ (it seems that this is SemA **1**) or it concerns also other semantemes.
   2.1. Sem-A **1** of ‘σ’ must belong/not belong to a specific class of entities or facts (see **2.4**: "be localized somewhere").
   2.2. Sem-A **1** of ‘σ’ must satisfy a complex set of conditions that concern other semantemes as well (see **2.5**: "causing something").

This is no more than a beginning of a fuller typology; at the time being, this is the first step to be taken. Note that proposed constraints on linguistically well-formed SemS are formally what are known as filter rules.

5 Concluding Remarks

Three questions will be touched upon in order to round up our exposition.

5.1 Minimality of SemS

The notion of well-formedness of SemSs is intimately related to the definition of SemS. It is impossible to discuss this definition here, but we will nevertheless indicate two properties of SemS essential from our perspective.

• The vocabulary of SemSs must be minimal (Apresjan, 1994 [1995: 468]): not all signifieds of full LUs of **L** are admitted as semantemes in a SemS. Thus, if the signifieds of two LUs L_1 and L_2 share a Sem-configuration ‘σ’, while the distinguishing parts (if any) contain only restrictions on the Sem-actants of ‘σ’, then this ‘σ’ is taken as a semanteme (rather than ‘L_1’ or ‘L_2’). For instance, the adjective HAZEL [X] ‘green-brown eyes [X]’ differs from GREEN-BROWN [X] only by a restriction on the SemA X; therefore, only the shared configuration ‘green-brown’ must be taken as a legitimate semanteme. The signifieds of the Russian causal prepositions IZ-ZA ‘because of’ and BLAGODARJA ‘thank to’ share the component ‘[Y] caused**1** by X’, and their differences are only restrictions imposed on their actants: (un)desirability of Y for the corresponding SemA of Y or for the Speaker; as a result, none of these signifieds is allowed to the SemS, the corresponding semanteme being ‘caused**1** by X’. In a similar vein, the Russian prepositions IZ, OT and PO share the Sem-component ‘[Y] directly caused**1** by X’ and are differentiated only by various

restrictions on their SemAs (Iordanskaja & Mel'čuk, 1996); all three are represented in the SemS by the Sem-configuration ‹directly caused**1**›.

• The composition of a specific SemS must also be minimal: not all elements of the extralinguistic situation represented by a SemS must (or even can) be reflected in it. Under lexicalization, the LUs selected can add semantic information and thus render the starting SemS richer. For instance, to express the semanteme ‹[Y] caused**1** by X› in Russian, one may have to choose between IZ-ZA and BLAGODARJA, and for this, one must know whether Y is desirable. It is impossible to require that any Russian starting SemS contain such an indication for everything that is caused**1**. Rather, the choice between IZ-ZA and BLAGODARJA triggers the search for this piece of information and in this way develops the "underspecified" starting SemS.

5.2 Extralinguistic and Linguistic Well-formedness of a SemS

The substantive well-formedness of a SemS can be characterized with respect either to extralinguistic reality or to the language under consideration.

• **Extralinguistic** well-formedness of a SemS corresponds to its **interpretability** in terms of extralinguistic reality (in the largest sense possible). This interpretability hinges on encyclopedic and pragmatic knowledge. For instance, the famous sentence *Colorless green ideas sleep furiously* has an encyclopedically ill-formed SemS: because of logical contradiction between semantemes, currently leading to absurdity (cf. ‹colorless green [ideas]›). From the linguistic viewpoint, however, its SemS is well-formed.

• **Linguistic** well-formedness of a SemS is its conformity to **language-specific constraints** on Sem-configurations. In other words, logical contradiction, tautology or any other type of absurdity do not automatically entail the linguistic ill-formedness of a given SemS.

A native speaker reacts to an utterance whose SemS is extralinguistically ill-formed by "What do you mean? This is nonsense." An utterance with a linguistically ill-formed SemS triggers a different type of reaction: "You don't say this like this; that's what you have to say."

Extralinguistic and linguistic well-formedness are logically independent properties of SemSs, so that four types of SemSs are possible.

1. Extralinguistically and linguistically well-formed SemSs underlie meaningful correct expressions: *Mary shot John*.

2. Extralinguistically well-formed, but linguistically ill-formed SemSs give rise to meaningful, but incorrect expressions: Rus. **Zmeja peremestilas´ iz nory* ‹The snake moved out of the hole› (⇒ *Zmeja vypolzla iz nory*).

3. Extralinguistically ill-formed, but linguistically well-formed SemSs produce meaningless expressions that can be linguistically correct or incorrect—at the level closer to the surface—as a function of the explicitness in expressing the contradiction. (By explicitness we mean using a separate full lexeme for a given meaning.)

> If the expression of a logical contradiction is not explicit, this contradiction gives rise to a linguistic anomaly (based on Apresjan, 1989 [1995: 624]).

Thus, in Russian:

— The synonymous phrases *kroxotnyj ogromnyj dom* ‹tiny huge house› and ?*kroxotnyj domišče* ‹tiny huge.house› have both extralinguistically ill-formed, but linguistically well-formed SemS; however, the first phrase is correct at the DSynt-level, while the second one is not, since the meaning ‹huge› is expressed in it by a derivational suffix **-išče**, i.e., not explicitly.

— The synonymous phrases *rovno desjat´ knig, xotja ja i ne znaju, skol´ko v točnosti* and **rovno knig desjat´* lit. ‹exactly books ten›, both meaning ‹exactly ten books, although I don't know precisely how many› are both extralinguistically ill-formed, but linguistically well-formed at the semantic level; however, the first phrase is correct at the DSynt-level, while the second one is not, since it expresses the meaning ‹although I don't know precisely how many› by a syntactic construction with inversion of the numeral and the quantified noun, i.e., not explicitly.

The sentence *Colorless green ideas sleep furiously* belongs here. In this connection, the following observation seems important. An extralinguistically ill-formed SemS normally features violations of some

constraints that the semantemes in this SemS impose on their SemAs: thus, ‹green› requires for its SemA **1** to denote a visible physical entity. However, when such a constraint, of a very general taxonomic character, is transgressed, the effect is an obvious absurdity expressible by a linguistically correct sentence. In sharp contrast, semantic constraints that control linguistic well-formedness of SemSs are much more specific and therefore much less obvious. Violating them produces the effect of a linguistic mistake.

4. Extralinguistically and linguistically ill-formed SemSs lead to meaningless incorrect expressions: Rus. **Zelënaja ideja peremestilas´ iz nory* ‹The green idea moved out of the hole› (no less absurd, but linguistically correct expression would be, for instance, *Zelënaja ideja vyletela/vypolzla/vyprygnula iz nory* ‹The green idea flew out/crawled out/jumped out of the hole›).

In this paper we have examined exclusively extralinguistically well-formed, but linguistically defective SemSs (Type 2 above).

5.3 Constraints on Sem-Configurations *vs.* Semantic Components in Lexicographic Definitions

Language-specific constraints on Sem-configurations cannot in principle be always reduced to particular semantic components in the decomposition of the semantemes under description. Such a constraint is of a different nature: it is itself not a *bona fide* Sem-configuration, but a requirement that a Sem-configuration must satisfy in order to be an actant or a Sem-governor of a given semanteme. Moreover, these constraints may be organized as complex Boolean formulas that cannot be easily included into a semantic decomposition; see, for instance, the restrictions on the use of the French semanteme ‹causer**1**› in (Kahane & Mel'čuk, 2006: 260). It follows that semantemes, although they are not signs, may have their own syntactics, which controls their combinability.

Acknowledgments

The draft of this paper has been read and commented upon by Ju. Apresjan, D. Beck, I. Boguslavskij, S. Kahane, J. Milićević, and A. Polguère; we thank all these colleagues and friends for their remarks and suggestions that have allowed us to significantly improve the presentation.

References

Apresjan Jurij. 1978. Jazykovaja anomalija i logičeskoe protivorečie. [1995: 598-621]

Apresjan Jurij. 1989. Tavtologičeskie i kontradiktornye anomalii. [1995: 622-628]

Apresjan Jurij. 1994. O jazyke tolkovanij i semantičeskix primitivax. [1995: 466-484]

Apresjan Jurij. 1995. *Izbrannye trudy. Tom II. Integral´noe opisanie jazyka i sistemnaja leksikografija*. Škola «Jazyki russkoj kul´tury», Moskva.

Apresjan Jurij. 2000. *Systematic Lexicography*. Oxford University Press, Oxford.

Iordanskaja Lidija & Igor Mel'čuk. 1996. K semantike russkix pričinnyx predlogov (*IZ-ZA ljubvi ~ OT ljubvi ~ IZ ljubvi ~ *S ljubvi ~ PO ljubvi*). *Moskovskij Lingvističeskij Žurnal*, 2: 162-211.

Iordanskaja Lidija & Igor Mel'čuk. 2004. The Meaning and Cooccurrence of Russian NEMNOGO 'a little.' In Ju. Apresjan (ed.), *Sokrovennye smysly: slovo, tekst, kul´tura. Sbornik statej v čest´ N. D. Arutjunovoj*, 112-127. Jazyki slavjanskoj kul´tury, Moskva.

Kahane Sylvain & Igor Mel'čuk. 2006. Les sémantèmes de causation en français. *LINX* [Revue de linguistique de l'Université Paris X Nanterre], nº 54, 247-292.

Malblanc, Alfred. 1961. *La stylistique comparée du français et de l'allemand.* Didier, Paris.

Talmy, Leonard. 2003. *Toward a Cognitive Semantics. Volume I. Typology and Process in Concept Structuring*. MIT Press, Cambridge, MA—London.

Lexical Functions vs. Inflectional Functions

Maarten Janssen
IULA / Barcelona
maarten.janssen@upf.edu

Abstract

Lexical Functions define structural relations between word-senses, based on semantic criteria. Inflectional functions, on the other hand, define derivational relations between words and their derived forms from a functional or semantic perspective. Inflectional functions are inspired upon lexical functions, and in many cases, the two types of functions define the same type of relations between the same pairs of words, but at a different level: the inflectional functions operate at the level of the lexical entry, the lexical functions at the level of the word-sense. This paper provides an overview of the similarities and differences between inflectional and lexical functions, and discuss the advantages and limitations of the use of inflectional functions in large-scale morphological databases.

1 Introduction

Lexical Functions (henceforth LFs) provide a way, amongst other things, to link several types of derived words to their morphological base in a structural, semantically oriented fashion (Žolkovskij and Mel'čuk, 1965) . For instance, the lexical function S_1 in (1) links the deverbal noun *walker* (the value of the LF) to its verbal base *walk* (the argument of the LF), while at the same time specifying the fact that the derived noun is in fact the agentive noun expressing *someone who walks*.

(1) S_1 (*walk*) = *walker*

That is not to say that the lexical function is intended to model morphological relations: the same relation of S_1 can hold in cases where there are two words with the right semantic relation between them, but where no morphological relation exists. An example is given in (2) , where *vendor* is the word for someone who sells, but it is not a word that is morphologically related to the verb sell *sell*. However, in the majority of S_1 cases the noun will be morphologically derived from the verb.

(2) S_1 (*sell*) = *vendor*

Given that lexical function are semantically-oriented relations, they relate word-senses rather than word-forms or words. For instance in the case of (2), the word *vendor* only relates to the verb *sell* in its basic meaning of 'giving or passing something in exchange for money'. Although you can paraphrase *John sells books* to *John is a vendor of books*, you can hardly say the same for the verb *sell* in the other meanings it is, as used in such sentences as *I sold the idea to my boss*, or *This CD sold five million copies*. Therefore, the argument in (2) should strictly speaking be interpreted as the first meaning (or lexie) of the word *sell*.

The fact that lexical functions operate on word-senses rather than words makes them even more fit for the representation of derivational relations: in general, derivations take a word as their input, but the meaning

of the derived word is (often) dependent on a specific meaning of the base word. For instance, the verb *problematize* means 'to turn something into a problem', but the word *problem* in this definition should be interpreted specifically as 'a difficulty that needs attention and thought' and not a problem in its meanings of a mathematical problem or someone who causes difficulty.

However appropriate lexical functions might be with respect to strictly derivational relations, they fare less well on a group of (derivational) forms that are partially inflectional in nature. These are the forms referred to as for instance *inherent inflections* (Booij, 1995) or *transpositional inflection* (Bauer, 2004). They are the forms that you do not find in the dictionary (or find in the dictionary as a *run-on* without definition), because they are assumed to be implicitly defined by the main article. Examples are (regular) diminutives like the Dutch *kindje* (small *kind* = child), deadjecival adverbs such as *roughly*, superlatives like the Spanish *bonísimo* (very *bueno* = good), etc. In this article, this class of forms will be loosely refered to as *inherent inflections*, though not necessarily with the same meaning as intended by Booij (1995; 2004).

One of the characteristics of inherent inflections is that, like inflection, they apply to entire words independent of the word-sense. Some of these forms are reflected by lexical functions, such as for instance the relation between a verb and its event nominal as in (3).

(3) S_0 (*produce*) = *production*

The noun *production* is the noun for the event related to the verb *produce*, independently of which of the several meanings of the words *produce* is taken into consideration. It is *the* nominal form of *produce* and not just a noun derived from it.

This means that in order to get the correct relation between each meaning of the verb *produce* and its respective event noun, there have to be as many instances of lexical functions between *produce* and *production* as there are meanings of the verb. Or in other words, the use of lexical functions for the relation between *produce* and *production* misses an important generalization, namely the fact that the relation is not restricted to a specific word meaning.

This lack of generalization was one of the motivations for the introduction of a notion of *inflectional functions* (henceforth IFs), which is a more morphologically oriented counterpart of lexical functions, operating at the level of words rather than at the level of word-senses (Janssen, 2005a). Inflectional functions have been subsequently implemented in a set of large-scale, full-form lexica called OSLIN (Open Source Lexical Information Network), starting with a lexicon for Portuguese, and currently with lexica for English, Spanish, and Catalan under development.

Given that inflectional functions are not merely a theoretical idea, but a practical solution implemented on complete lexica, the use of IFs unavoidably ran into conceptual and pragmatic problems. This article will fist explain the notion of inflectional functions and its implementation in the OSLIN databases, and then discuss the advantages and limitations of inflectional functions in detail.

2 Inflectional Functions

Inflectional functions define relations between lexical entries and their derived forms. They therefore in principle model derivations, but are called inflectional because they deal with the forms that are midway between derivation and inflection, the forms of what you might called the *extended paradigm*. IFs do not model the process of word-formation: they do not indicate the type of process that led to the creation of the one from the other, but they define a semantic or functional relation between words independently of the actual process that was involved in the morphological derivation. An example of an inflectional function is given in (4), where **s0v** defines a relation between a verb and its related event nouns.

(4) **s0v** (*descend*) = *descent*

Deverbal event nouns are not the only inherent inflections modeled by IFs: other IFs are the female forms of animate nouns (not very productive in English, but *actor* - *actress* would be an example), diminutive forms of nouns, comparative adjectives, etc.

Inflectional functions are mid-way between morphological relations and semantic relations, in the sense that they define relations with both morphological and semantic characteristics. For instance, in order for there to be an **s0v** relation between two words X and Y, the following conditions have to be met:

(a) X has to be a verb, and Y has to be a noun

(b) Y has to be morphologically related to X

(c) Y has to be an event noun referring to the abstract event expressed by the verb X

(d) Y has to apply to all meaning of the verb X

Because of the semantic requirement (c), inflectional functions do not provide purely morphological information. Although the noun *ignorance* is a deverbal noun, derived in the same was from the verb *ignore* in the same way as the word *acceptance* is derived from the verb *accept*, there is a **s0v** relation between the latter two, but not between the former. Synchronically there is no direct semantic relation between *ignorance* and *ignore*, although probably there originally was. Because of the morphological requirement (b), IFs are not really semantic in the sense that LFs are (they are not used for cases like *vendor*). And because of the lack of an indication of the morphological process, IFs do not provide full morphological information since they do not specify *how* the noun is derived from the verb, but merely require that it is.

It is mainly requirement (d) that makes inflectional functions usable mainly for inherent inflections, and not for just any type of derivational relation. Inherent inflections are derivational relations that are largely inflectional in nature in the sense that the derived word can be seen as a context-dependent word-form of the word it is derived from. For instance, in the case of the relation between *descend* and *descent* in (4), the noun can be seen as the nominal form of the verb, that is the say, the form of the verb that should be used in denominal constructions. For instance, the two sentences in (5) express the same content, except that the second sentence uses a light-verb construction. Because of the light-verb construction, the use of the verb *descend* in its nominal form is required, and therefore the adverb *rapidly* has to be used in its (base) adjectival form. You could therefore say that the shift for a verb to a noun is a case of (transpositional) agreement.

(5) He descended rapidly. ⇒ He made a rapid descent.

Inherent inflections also behave like regular inflections in the sense that the form the event noun takes is not dependent on the meaning that the verb is used in: the form *sold* simply is the past tense of the verb *sell*, and not the past tense of the verb *sell* in a specific meaning. It cannot be the case that when we are talking about selling of ideas for instance, the past tense takes a different form, say *selled* (with some exceptions which will be discussed later). And neither can the nominal form be anything different from *descent* when we use the verb in for instance its meaning of 'the arrival of many people at the same time'.

Although the inherent inflection has to be the same for each meaning of the base word, it does not have to be unique. There can be various alternative deverbal event nouns for the same verb. For instance, the event noun for the Portuguese verb *manifestar* (to manifest) is *manifestação*, but also *manifesto*. This is comparable to the situation for verbal inflection: the past tense of *blow* is either the irregular *blew* or the modern day regularization *blowed*. But both past tenses are correct for each meaning of the verb *blow*, and so are both *manifestação* and *manifesto* correct deverbal nouns for each use of the verb *manifestar*. There can be differences in register between the variants, and there can even be a preference for a specific form related to a specific meaning, in the sense that *blowed* is a colloquial form, mostly used for colloquial meanings of the verb, but all variants should be correct for each meaning.

2.1 Words and Inflectional Functions

Inflectional Functions operate on the level of *words*. What exactly a word is in this context is best understood by looking at inflection proper. The past-tense *rung* does not belong to the sequence of letters *ring*, but to the verb *(to) ring*, since the noun *ring* does not have a past tense. It does not even belong to "any" verb *ring*, but to one of the homonyms of *ring* (to call; to sound a bell), since there is another verb *ring* (to put a ring around the foot of a bird) which has the past tense *ringed*. And the past tense also does not really belong to the abstract notion of a word but is more anchored in orthography: the Portuguese word for *wet* can be written as either *húmido* or *úmido* and the two forms are typically said to be different ways of writing the same word. But the female form *húmida* only belongs to the first form, and not to the second. Inflectional paradigms belong to something that is somewhere in the middle: the lexical entry, the headword as it appears in the dictionary, or to what in MTT is called the *vocable* (Mel'cuk et al., 1995).

The same holds for the extended paradigms defined by the inflectional functions. When there are different orthographic realizations of a verb, there can be different deverbal nouns for each of them: the Portuguese noun *doiramento* is the **s0v** of *doirar* (to gild) and not of its orthographic variant *dourar*. And when a noun is homonymous, it can have different female forms for each homonym: the Catalan noun *croat* has a female form *croata* when it indicates someone from Croatia, but a *croada* when it is a female crusader. But all the different word-senses (or lexies) related to a polysemous adjective have to share the same superlative form(s). So the argument of an inflectional function is a lexical entry, just like a in the case of inflection proper[1].

Although (extended) inflectional paradigms belong to lexical entries, not all the inflectional forms have to be realizable in each word-sense. There is only one noun *water* in the English language, and its plural is *waters*. But most commonly, *water* is used as a mass noun and does not have a plural. To avoid having to resort to defining homonymous entries for all words that can be used as mass nouns or count nouns (which given the *universal grinding* mechanism would be most nouns), or having to define inflection at the level of word-senses, one has to conclude that inflectional paradigms can be defective in only some of the acceptions of a lexical entry: *waters* is the plural for *water* in all senses, but it is always realized. The same holds for inherent inflections: the Portuguese word *amarela* is the female form of the noun *amarelo* (yellow), but can be used only when the word *amarelo* denotes an animate object (a pale person). We call this phenomenon *partial defectivity*.

Since the argument of an IF is not a word-form, neither is its value: the Catalan word *peixatera* (fishwife) is the female form of the word *peixater* (fishmonger). But *peixatera* is a lexical entry in its citation form, and has two inflected forms *peixatera* and *peixateres*. The value of an IF is not a lexical entry either: only in its basic meaning is the word *construction* the **s0v** of *construe*; when used for a building, it is not a deverbal event noun. Therefore, the value of an IF is in principle a word-sense, as indicated in figure 1.

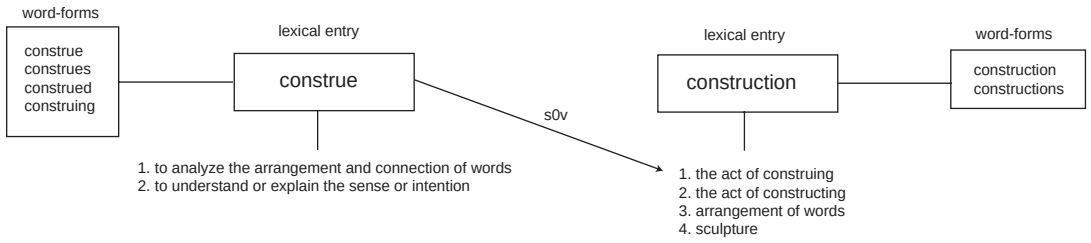


Figure 1: Inflectional Function for *construction*

Inflectional functions are explicitly intended to model forms that are mid-way between derivation and

[1]The fact that the boundary between polysemy and homonymy is not absolutely clear-cut can lead to problems, as will be discussed in 3.2.

inflection, and model them in such a way that they can be interpreted either way depending on the theory/application they are used with. When considering the relation **female**, it can be taken to define morphologically derived forms, or it can be taken to define (additional) inflected forms for the base noun, as is often done for the Romance languages. But since the value of the IF is not a word-form, interpreting them as inflectional has to be done indirectly: if we want to take **s0v** as an inflectional relation (see figure 1), the word-forms (*construction* and *constructions*) of the lexical entry (*construction*) that the value of the s0v ('the act of construing') belongs to should be taken as inflectional forms of the lexical entry (*construe*) that is the argument of the lexical function. So indirectly, *construction* and *constructions* would be (inherently inflectional) forms of the lexical entry *construe* if **s0v** is taken to define an inflectional relation. Along the same lines, *peixatera* and *peixateres* would end up as additional inflectional functions of *peixater*.

2.2 The Open Source Lexical Information Network

The Open Sources Lexical Information Network (OSLIN) is a framework for relation-based, large-scale, full-form lexica (Janssen, 2005b). The first lexicon developed in the framework was a lexicon for Portuguese called *MorDebe*, built at the ILTEC institute in Lisbon. MorDebe contains around 135.000 lemmas and just under 1,5 million word-forms. Currently, lexica for Catalan, Spanish, and English are under development at the IULA institute in Barcelona, each of which currently contains over 100.000 lemmas. All these lexica are in a first instance meta-lexicographic, providing a more structured representation of the nomenclature of existing dictionaries. But the lexica have a mechanism for maintenance, and are integrated with a semi-automatic neologism detection program called NeoTrack (Janssen, 2008), meaning that they are gradually being expanded with additional lemmas. All the OSLIN lexica can be consulted online via *oslin.org*.

OSLIN is a full-form lexicon model that not merely specifies all inflectional forms explicitly, but also attempts to list all the inherent inflections of the extended paradigms. For the most extensively developed lexicon, the Portuguese MorDebe, all dictionarized deverbal event nouns were collected, as were all the dictionarized nominal gender cases (*gato - gata*) and all the deadjectival quality nouns. Apart from these extensively explored relations, a number of other lexicalized relations was collected as well, including diminutive forms and adjectival superlatives. All these relations are modeled in terms of inflectional functions. The relations used, as well as the number of relations stored, are listed in table 1. As a special case, also the *gentiles* were collected extensively, which are not relations between lexical entries and derivatives, but between toponyms and the nominal or adjectival form associated with it.

s0v	7347	event noun	acertar	-	acerto
female	8124	nominal gender	actor	-	actriz
adv0	1974	adverb of mood	abjecto	-	abjectamente
a2v	3716	participial adjective	abalar	-	abalado
able2	239	-able adjective	absorver	-	absorvível
dim	265	diminutive	animal	-	animalzinho
aum	79	augmentative	monte	-	montão
max	265	adjectival superlative	magro	-	macérrimo
genta	1561	gentile	Aachen	-	aacheniano

Table 1: Inherent inflections in MorDebe

Inherent inflections are in general rather regularly formed, although there are some irregular forms. But for many of the inflectional functional relations, there are various ways for deriving the same form. For instance, deverbal event nouns can be regularly formed in Portuguese with the suffixes *-ção, -mento, -dura, -dela*, and *-o*. For some verbs, all of these are used. For instance, for the verb *abafar* (to choke), there are the forms *abafação, abafamento, abafadura, abafadela, abafo*, and *abafeira*. But in most cases, only one

of the potential forms is actually used, and the others are considered incorrect. The correct event noun for *abolar* (to do away with) is *aboladura*, for *abdicar* (to abdicate) it is *abdicação*, for *abocar* (to put into your mouth) *abocamento*, whereas other forms such as **abocação* and **abolamento* are not correct. Which form(s) is the correct one is something that has to be learned/defined for each individual verb. Therefore, using derivational rules for these kind of relations helps to predict potential forms, but does not specify the actual lexical items used. The use of inflecitonal functions to model these relations systematically has proven to be a powerful way of structurally treating this strongly lexicalized type of information.

2.3 Comparison between LF and IF

Lexical functions and inflectional functions are fundamentally different animals: the arguments of IFs are lexical entries, the arguments of LFs are word-senses (lexies). IFs are morphological in nature, LFs are not. And LFs can be used to specify collocational relations, whereas IFs cannot. However, for many of the IFs, there is a corresponding LF, as shown in table 2. The first items on the list are standard LF, the ones with a star are the LF proposed in the extension of the DECIDE project (Grefenstette et al., 1996).

s0v	event noun	S_0	0-th role noun
s0a	quality noun	S_0	0-th role noun
a2v	participial adjective	A_2	2nd role adjective
able2	-able adjective	Able*	Adjective of possibility
female	gender noun	Female*	Animate nominal
max	adjectival superlative	Max*	Maximal degree

Table 2: Inherent inflections - Lexical Functions

Because of the difference in nature, there is no direct mapping between LFs and IFs: not every S_0 will have a corresponding s0v. Not only because the S_0 relation is also used for deadjectival nouns, but also because IFs require a morphological link between the two elements of the relation whereas the LFs do not. Therefore, there is a Female LF relation between the Dutch words *kat* (cat) and *poes* (she-cat), but there is no **female** IF relation between them because *poes* is not morphologically derived from *kat*. And similarly, there is a S_0 relation between *steal* and *theft*, but there is no corresponding **s0v** because there is no morphological relation between the two words.

But although S_0 relations do not automatically lead to **s0v** relations, for every IF for which there is a corresponding LF the first implies the second. So the fact that there is a **s0v** relation between the verb *coincide* and *coincidence*, implies that there will be a S_0 relation for each of the word-senses of the verb *coincide* (although there are more word-senses for the noun *coincidence* that are not related to *coincide*, and there might be cases of partial defectivity). In that sense, the OSLIN lexica provide large-scale repositories of lexical functions, encoded in an efficient way.

3 Practical and Theoretical Issues

The advantages of using Inflectional Functions over Lexical Functions in a full-fledged lexicon system like OSLIN are easy to point out. First of all, OSLIN at this moment only contains word-forms and lexical entries (and proper names). Although it is the idea that in the future, a level of word-senses will be added, at this point in time that is still not the case. In the absence of word-senses, it is impossible to define LFs, since LFs by definition require word-senses.

But even if there were word-senses in OSLIN at this time, modeling the types of relations currently defined by IFs in terms of LFs would not only be inefficient, but also theoretically inaccurate. Consider a relation which could be expressed by lexical functions: diminutives. In Catalan, there are various highly polysemous nouns, such as for instance the word *mà* (hand), for which the academy dictionary (DIEC2)

lists 17 main senses and a whole range of sub-senses. In all of these meaning, the word *mà* can be expressed in diminutive form, and in each of the senses, the diminutive form is *maneta*. Listing an extensive list of relations between *mà* and *maneta* in every sense would not only create a redundant amount of relations, it would also incorrectly suggest that it is something specific for each individual meaning of *mà* that its diminutive form is *maneta*, whereas in fact it is a entry-based property of the word *mà*.

However, despite the obvious advantages, the application of inflectional functions in OSLIN raised several complications. Some of the most important ones will be discussed in this chapter. The first problem, that of the failure to distinguish between alternative forms, is not specific for inflectional functions, but equally affects lexical functions. The other problems however, are more particular for inflectional functions.

3.1 Distinguishing Alternatives

As mentioned in the previous chapter, there are often various alternative forms for the same inherent inflection. For instance, for the verb *lavar* (wash) there are 6 different event nouns mentioned in the dictionary: *lavação, lavadura, lava, lavadela, lavagem*, and *lavamento*. In principle, all these words express the same concept: the act of washing (or cleaning). But not all these words express that meaning in exactly the same way, and not all words are usable in the same context. What exactly is the difference between them is difficult to make explicit, but the word *lava* is uncommon, and is unlikely to be used in every-day expressions. The word *lavadela* has a diminutive ring to in, and is mostly applied to light washings. Where *lavagem* is used for things that have more to do with the effect of washing, *lavemento* is more associated with the result of washing (although the difference between those is far from clear).

Inflectional functions do not provide any means to distinguish between the different ways of expressing the different event nouns for *lavar*, or distinguish between any of the other alternative forms of inherent inflections. This failure to distinguish between different alternatives, or even indicate which is the most frequently used or common one, is a serious drawback in the current presentation of the inherent inflections in the (Portuguese) lexicon. And alternative inherent inflections are not rare: for instance for the deverbal nouns, there is more than one alternative for about 30% of all the verbs.

Notice that this is not a problem that is specific for the inflectional functions: since all of the event nouns are usable for each of the meanings of *lavar*, and definitely all of them with the primary meaning of *lavar*, the relation S_0 between the word-senses of *lavar* and the various event nouns would equally fail to distinguish the subtle differences between them. In fact, the problem is even slightly bigger for the lexical functions, since in the case of LFs, also the non-morphologically related event nouns are taken into account, and the lexical functions are not capable of distinguishing between *stealing* and *theft* as the S_0 for *steal*. It is an indication that although relational models like IFs and LFs partially define the meaning of the words they relate to, they do not define the meaning and use in full.

3.2 Stacking Failures

Inherent inflections can in principle be stacked. As shown in example (6), the female form and the diminutive form can be stacked to derive the female diminutive form *gatinha* (little she-cat). And as shown in (7), there is even more than one way to obtain the same result by stacking the diminutive and the female in the reverse order.

(6) **female** (*gato*) = *gata*, **dim** (*gata*) = *gatinha* ⇒ **dim**(**female** (*gato*)) = *gatinha*

(7) **dim** (*gato*) = *gatinho*, **female** (*gatinho*) = *gatinha* ⇒ **female**(**dim** (*gato*)) = *gatinha*

In this particular case, the meaning of the two is identical, although this is not alway the case: there are such cases as the negation of the -able form of *do* and the -able form of the negation of *do*. In both cases, the correct result is *undoable*, but depending on the order of application, it is either something that cannot be done, or something that can be undone, which is hardly the same thing.

Although in principle stacking is a positive feature of inflectional functions, in certain cases it leads to undesirable results, because of two features of inflectional functions. The first feature involved is the fact that with respect to their inherent inflections, paradigms can be partially defective as explained in 2. The second feature involved is the fact that although inflectional functions take lexical entries as their arguments, the *value* of an inflectional function is in principle a word-sense and not a word: the word *llibret* in Catalan is the diminutive of *llibre* (book), but the diminutive form has obtained the lexicalized meaning of the little booklets used in the opera (libretto). As a pragmatic solution to this problem on the OSLIN website(s), in those cases, the entry for *llibret* does not display that it is the "diminutive form of *llibre*", but that it is "*also* the diminutive of *llibre*", indicating that it has other meanings as well.

When these two features apply to the same word, the situation becomes rather complicated. The word *ilha* (island) in Portuguese has several diminutives forms, one of them being *ilheu*. But apart from being the word for a little island, the word *ilheu* also means somebody living on the island, or an islander. Now the female form of *ilheu* in its meaning of an islander is *ilhoa*. Therefore, the definition for the word *ilheu* says that it is (also) the diminutive of *ilha*, and that it has a female form *ilhoa*. But given that the are no female little islands, the female forms does not apply in the case of the primary meaning of *ilheu*. Although this technically speaking does not say anything incorrect, it really stretches the limits of what can be done with inflectional functions. For those cases, a more word-sense driven approach like lexical functions would be much less problematic.

It should be noted that however problematic this stacking problem is, it is a practical, and not a theoretical issue: since the argument of an inflectional function is a lexical entry, and the value a word-meaning, the value a one lexical function cannot be taken as the argument of a next one. That means that strictly speaking, inflectional functions cannot be stacked as in (6). But in the practical use of inflectional function, cases such as *ilha-ilheu-ilhoa* present a serious problem for a coherent treatment.

3.3 Sense-Specific Exceptions

One of the crucial features of inherent inflections is that it is a lexical entry-level phenomenon. This implies amongst other things that all event nouns that are specific to a given word-sense should not be considered inherent inflections. Up to a point, this situation is comparable with the situation of regular inflection. The aforementioned meaning dependent past tense *ringed* versus *rang/rung* can only exist because the word *ring* is homonymous. Within the OSLIN framework, the existence of multiple inflectional paradigms is even taken as a criterion for homonymy. Since inflection is taken as a definitional criterion for the identity of words, even cases like *hang* are considered homonymous: it has either *hung* as its past tense, or *hanged* in its meaning of *killed by hanging*. This despite the fact that the latter is clearly both etymologically and semantically related to the general meaning of hang. Modeling the inflection correctly without assuming there to be two lexical entries for *hang* is extremely complicated, both from a practical and from a theoretical perspective.

For inherent inflection, basically the same holds: if a word has two different female gender nouns, as in the case of the *croat* in section 2.1, then the word has to be considered homonymous. However, given that inherent inflections are not really inflectional, there are several possible solutions when there are two meaning-specific inherent inflections for a given word. Consider the French word *fille* (girl; daughter)[2]. Most French dictionaries list a single entry for both meanings of the word, considering it to be a case of polysemy rather than homonymy. However, the diminutive form *fillette* can only be a small girl, and not a small daughter. This means that *fillette* can not be considered a meaning-independent inherent inflection of *fille*. There are three different ways to solve this problem.

The first option is to say that partially because of the different diminutive forms, the word *fille* should be considered homonymous, contrary to how dictionaries treat them. There are independent reasons for doing

[2]Example provided by an anonymous reviewer

so (Polguère, 2008), but the different extended paradigms could be taken as a strong indication that the word is synchronically really homonymous rather than polysemous. In that case, the duplication of the lexical entry will solve the problem since the diminutive will apply to all meanings of one of the entries.

The second option is to say that *fille* is partially defective: it is not that there is another diminutive for the meaning of 'daughter' that is correct, it is that in its meaning of a daughter, the word does not have a diminutive. Since there is no diminutive for the French words *fils* (son), *père* (father), or *mère* (mother) either, this could simply indicate that daughter is not of the right type to take a diminutive. That would mean that *fille* does not have a diminutive in its meaning of 'daughter' in the same way as *water* does not have a plural as a mass noun.

The third option is to say that the word *fillette* in this case is not an inherent inflection, but rather a meaning-dependent derivational form that happens to be formed as a regular diminutive. It is not transparently *the* diminutive form of *fille*, just a word for a small girl. In that case, there will not be an inflectional function between the two words, and the meaning-specific relation will have to be treated with, for instance, lexical functions.

Which of these three solutions is the most correct one is something that has to be considered for each individual case. In the case of *fillette*, the first of these options seems to be the most reasonable, but all three solutions will resolve the problem of seemingly meaning-specific inflectional functions. But it should be noted that for inflection proper and the clear cases of inherent inflections, cases of potential meaning-specific inherent inflections are rare.

3.4 Marginally Inflectional Forms

In the application of inflectional functions to the entire lexicon, it is obvious that there are many marginal cases. For instance, when creating a full list of all **s0v** cases, there are many examples in which it is debatable whether the noun in question is really an **s0v** deverbal noun. The first major reason for doubts is that it is sometimes not clear whether there is a morphological link between the verb and the noun. In Portuguese, many deverbal nouns are direct adaptations from Latin, and not (or not only) synchronically derived: the word *confusão* (confusion) comes from Latin, and there is no synchronic operation to derive it from the verb *confundir*. The second major reason for doubts is that there are many cases in which it is unclear if the noun (still) expresses the abstract event related to the verb, or whether it can do so for all meanings of the verb. These problems can only be solved on a case-by-case basis, using well-defined criteria to render a consistent database of deverbal nouns.

Note, however, that there are also classes of derivational forms that structurally have a marginal character. Deadjectival quality nouns in many senses behave like deverbal event nouns. The nominal form applies generally to all meaning of the adjective - a polysemous adjective like *happy* has the nominal form *happiness* in every meaning of the word. And the nominal form hardly expresses anything else than the abstract quality related to the adjective, expressed in a nominalized way. This is examplified by the two sentences in (8) which are largely synonymous.

(8) He is very happy. ⇒ He has a lot of happiness.

However, different from event nouns, the quality nouns are quite regularly related to a specific meaning of the adjective, or a range of meanings of the adjective. For instance, the Portuguese adjective *fino* (fine) has a number of meaning, and a number of related quality nouns. But which quality noun is the correct one often depends on the meaning of the adjective. The noun *finura* relates to *fino* in its meaning of 'delicate', whereas the word *fineza* means 'thin-ness'. And similarly, with respect to the adjective *bravo* (angry; brave), *bravura* means 'braveness', and *braveza* is the act of being angry (Correia, 1999).

A possible conclusion, and the current pragmatic solution in OSLIN, is that deadjectival quality nouns are not inherently inflectional, but really derivational like agentive nouns are, and should therefore not be modeled by means of inherent inflections. However, for most adjectives, the derived quality noun is not

meaning specific, and the noun reflects the abstract property related to any of the meanings of the adjective. Therefore, treating quality nouns for polysemous adjectives with lexical functions fails to take the general character of the relation between argument and value into account, in much the same way as it does for deverbal event nouns. This means that quality nouns are on the one hand too much meaning independent to be treated with purely word-sense based lexical functions, but on the other hand to meaning specific to be treated by the purely lexical entry oriented inflectional functions. And quality nouns are not unique in this respect: the same holds for instance for the agentive nouns as well.

4 Conclusion

As shown in this paper, inflectional functions provide an efficient way of modeling inflection-like derivational relations such as deverbal event nouns, female nouns, diminutives, and superlatives. Several of these relations are currently modeled by lexical functions, and inflectional functions provide an entry-level treatment of these relations rather than a word-sense based treatment, which is not only more general, but also matches the linguistic reality closer. Inflectional functions imply lexical functions, which means that the OSLIN databases in which inflectional functions are applied at large scale, indirectly provide large repositories of lexical functions.

But as also shown in this paper, there are several cases in which the application of inflectional functions is problematic, running into the boundaries of what can feasibly be expressed in terms of such entry-wide relations. However, despite these limitation, inflectional functions still provide a more efficient and manageable way of representing lexical relations in large-scale full-form lexica providing information on the extended paradigm of words.

References

Bauer, Laurie. 2004. The function of word-formation and the inflection-derivation distinction. In Aertsen, Hannay, and Lyall, editors, *Words and their Places. A Festschrift for J. Lachlan Mackenzie.* Vrije Universiteit, Amsterdam.

Booij, Geert. 1995. Inherent versus contextual inflection and the split morphology hypothesis. In Booij and van Marle, editors, *Yearbook of Morphology 1995*. Kluwer, Dordrecht.

Booij, Geert. 2004. *The Grammar of Words*. Oxford University Press, Oxford.

Correia, Margarita. 1999. *A denominação das qualidades: contributo para a compreensão da estrutura do léxico português*. Ph.D. thesis, Universidade de Lisboa.

Grefenstette, G., U. Heid, B. M. Schulze, T. Fontenelle, and C. Gerardy. 1996. The DECIDE project: Multilingual collocation extraction. In *Papers submitted to the Seventh EURALEX International Congress on Lexicography*, Göteborg, Sweden.

Janssen, Maarten. 2005a. Between inflection and derivation. In *East-West Encounter: Second International Conference on Meaning ⇔ Text Theory*, Moscow, Russia.

Janssen, Maarten. 2005b. Open source lexical information network. In *Third International Workshop on Generative Approaches to the Lexicon*, Geneva, Switserland.

Janssen, Maarten. 2008. Neotrack: Une analyseur de nologismes en ligne. In *Actes de CINEO 2008: Actes del I Congrès Internacional de Neologia de les Llengües Romàniques*, Barcelona, Spain.

Mel'cuk, Igor, André Clas, and Alain Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*. Duculot, Louvain-la-Neuve.

Polguère, Alain. 2008. *Lexicologie et sémantique lexicale. Notions fondamentales*. Les Presses de lUniversité de Montréal, Montréal.

Žolkovskij, A. and I. Mel'čuk. 1965. O vozmožnom metode i instrumentax semantičeskogo sinteza. *Naučno-texničeskaja informacija*, 5:23–28.

Defining the Deep Syntactic Structure: How the signifying units combine

Sylvain Kahane
Modyco, Université Paris 10 & CNRS
Alpage, INRIA & Université Paris 7
sylvain@kahane.fr

Abstract

Considering that the DSyntS has been described in detail but never really defined, we propose to define it as the structure recording how the signifying units combine when an utterance is produced. Signifying units are defined as the indivisible choices made by the speaker as (s)he produces an utterance; they include full lexical units but also grammatical, derivational and constructional units. The traditional DSyntS is a simplified view on the structure we define here. We show that many choices made for the traditional DSyntS are arbitrary and sometimes irrelevant. We also discuss the status of the DSyntS in a complete linguistic model.

1 Introduction

The Deep Syntactic Structure [henceforth DSyntS] is probably the mostemblematic structure in Meaning-Text Theory [MTT]. Although the DSyntS is used in many studies, I think it has never been correctly defined and many choices remain obscure and insufficiently argued for. This problem was stated in two of my previous communications to the Meaning-Text Conference (Kahane 2003a, 2007). A recent study of Mel'čuk & Wanner (to appear) focusing on the grammatical elements in the DSyntS and with which I have numerous disagreements motivated me to make a new attempt to define the DSyntS.

In fact, all the presentations of the DSyntS proposed by Mel'čuk (1974, 1988, to appear) pose two problems. First, the DSyntS is presented as one of the seven levels of linguistic representations foreseen in a Meaning-Text model [MTM], intermediate between the Semantic and the Surface Syntactic levels; it does not have a characterization of its own, independent of the whole model. Secondly, the DSyntS is described rather than defined or characterized. In other words, there is absolutely no way to falsify anything, which is very problematic from the scientific point of view. I think this is why, though the DSyntS has often been recognized as an interesting level of representation, it has not achieved the success it deserves.

This presentation tries to solve both problems, firstly by giving a characterization of the DSyntS as independent as possible from the rest of the theory and secondly by exploring the various possible representations compatible with this characterization and showing that the traditional presentation of the DSyntS is not necessarily the most relevant.

The main purpose of Section 2 is to show that the DSyntS expresses the organization of a natural class of linguistic units, which we call the signifying units and which can be defined independently of the rest of the theory. Section 3 examines exactly how the signifying units are organized within an utterance as well as the nature of the resulting structure, the DSyntS. Several difficulties are tackled in these two sections: grammatical units, systematic polysemy links, free government patterns and communicative constructions. Given the definition of the DSyntS given here, Section 4 explores its status from the viewpoint of a monolingual MTM as well as from the viewpoint of paraphrasing and translation.

2 The DSynt units: the signifying units

The definition of the DSyntS we propose is entirely based on the notion of *signifying unit* [SU], which we will try to define here. The term is a translation of Fr. *unité signifying* introduced by Saussure (1916) and defined in the sense we use here by Martinet (1960) (see also Ducrot 1995). It is also the *minimal semantic constituent* of Cruse (1986:25).

2.1 Paradigm of choice

Our definition of SUs is based on the notion of *choice* introduced by Martinet (1960:26) in a chapter entitled *Every unit presupposes a choice* [I translate]: "Let us consider an utterance like *this is a good beer*. […] If we are able to say something about the combinatorial latitudes of *good*, it is because this segment of the utterance has been recognized as a particular unit distinct from *a* and *beer*. To reach this result, one must have noticed that *good*, in this context, corresponded to a specific choice between some other possible adjectival modifiers; the comparison with other utterances has shown that in the contexts where *good* appears one also find *excellent, bad*, etc. This indicates that the speaker, more or less consciously, moved apart all the competitors which could have appeared between *a* and *beer* and which was not considered appropriate in this case. Saying of the hearer that (s)he understands English implies that (s)he identifies by experience the successive choices the speaker must have done, that (s)he recognizes *good* as a distinct choice from *a* and *beer*, and that it is not excluded that the choice of *good* rather than *bad* influences her/his behavior."

Let us take the utterance *Peter bought an eggplant* as an example.[1] In this utterance *egg* and *plant* are identified as well-known lexemes of English, but none of them results of a choice: neither *egg* has been chosen by opposition with *ball* or *testicle*, nor *plant*, by opposition with *fruit* or *vegetable*. It is *eggplant* in its entirety which has been chosen by opposition with *carrot, French bean* or *cauliflower*. Thus *eggplant*, *carrot, French bean* or *cauliflower* form a *paradigm of choice* or *system of oppositions*, where each of these choices is indivisible. We call *signifying unit* [SU] any linguistic sign that presupposes an indivisible choice by the speaker.[2]

In Kahane (2007), I gave the example of the French utterance *La moutarde me monte au nez* (lit. The mustard goes to.me up to the nose, 'I feel anger welling up in me'), where 4 choices are made by the speaker and there are a corresponding number of signifying units: the phraseme ⌜LA MOUTARDE MONTER AU NEZ⌝, the pronoun MOI (in its clitic form *me*) which is the single actant of this phraseme, the present tense and the declarative construction.

We distinguish four types of SUs: lexical units [LUs], including idioms and lexical functions, grammatical units [GUs], derivational units [DUs], including systematic polysemy links, and constructional units [CUs].

Mel'čuk does not distinguish the nature of the Deep-Syntactic [DSynt] units from the Surface-Syntactic [SSynt] ones; of course, there are some units which cannot appear in DSyntSs like agreement grammemes or others which cannot appear in SSyntSs like phrasemes, but most of them can appear both at the DSynt and SSynt levels. I think that DSynt and SSynt units are different in nature. When we speak about LUs we only consider what Mel'čuk calls *full* lexical units; I do not think that there are others and I think that the units of the SSyntS must be different from LUs.

All the notions introduced in this subsection will be clarified in the next subsections.

---

[1] The recommended spelling of *eggplant* is in one word, but there are much more spellings *egg plant* than *eggplant* on the web. For our discussion, however, it does not make any difference whether it is one or two words.

[2] The term *choice* (*by the speaker*) evokes modeling a cognitive process during the enunciation. We do not exclude that it could be relevant, but it is not our task here. We put ourselves in the framework of the distributional analysis and our notion of (paradigm of) choice is uniquely based on substitution (see next section).

2.2 Compositionality and collocations

It is tempting to use the notion of (semantic) compositionality to define SUs, but it poses some problems we will consider here.

A combination of linguistic signs AB is generally said to be *compositional* if the meaning of AB is constructed by a regular combination of the meanings of A and B. This definition would only be operational if we were able to represent and compute the meanings of A, B, and AB and to prove that 'AB' is a simple combination of 'A' and 'B'. We can nevertheless agree that some combinations are compositional without doing that. The question is particularly challenging for morphemes that can never be used autonomously like *-er* in *killer*. If *killer* is compositional, it is because in *killer*, *kill* can be substituted by *sell* or *run* and the contribution of *-er* is about the same in every case (and *kill* has the same contribution as in its verbal use). This can be formalized by the following definition.

> We say that the linguistic sign A can be *properly substituted* by A' in the combination AB if: 1) A and A' are mutually exclusive, that is, B cannot simultaneously combine with A and A', and 2) the interpretation of B is not modified by the substitution, that is, the semantic ratio of 'A'B' on 'A'' is the same as the semantic ratio of 'AB' on 'A'; in particular, if A' and A are synonymous, A'B and AB must be synonymous. (See Cruse 1986 for a similar definition).
>
> We say that a linguistic sign A can be *freely substituted* in the combination AB if 1) the set of elements that can be properly substituted for A is rather regular and notably can be deduced from the sets of elements that can be substituted for A in other combinations and 2) in the set of elements that can be properly substituted for A in the combination AB, there is an important proportion of elements that have a distribution similar to A.[3]
>
> We say that AB is a *free combination* if both A and B can be freely substituted in the combination AB.

In the sentence *The boy cried*, each of the four segments *the, boy, cry-* and *-ed* can be freely substituted, as well as the segments *the boy* and *cried*.[4]

We can now get back to the characterization of SUs. It must be first noted that every sign that can be decomposed into a free combination of signs is *not* an SU. But the converse is not true: Some combinations that are not free can be described as a combination of SUs. Such a restricted combination cannot result from the combination of two independent choices. But it is possible that one choice has been made freely and that the second choice has been made according to the first one. This is exactly the case of collocations.[5] In a collocation like *heavy rain*, the base RAIN is freely chosen, while the choice of the collocate HEAVY (whose substitution by a synonym like BIG gives an odd sentence) is constrained by the base. Such constrained choices are modeled in MTT by lexical functions [LFs]. A LF applies to a LU (the base) and gives all the possible expressions of a given meaning in the context of this base (the collocates). For instance, the LF Magn maps the meaning of intensification onto HEAVY when it applies to RAIN. A LF can be considered as a generalized SUs whose signifier varies with the context and can be put in a DSyntS in place of one of its value, as usually done in MTT.

---

[3] This definition could sound vague. But the notion of free combination is a gradual notion: The more regular is the set of elements that can be substituted for A and the more important is the proportion of elements having the same distribution as A, the freer is the substitution.

[4] Indeed the set of elements that can be properly substituted for *boy* in this context is the set of human nouns, which can be found in many other contexts. The set of elements substitutable for *-ed* consists of only two other elements, *-s* and a zero suffix, but this set combines with many other verbs and, modulo allomorphs, the three morphemes have the same distribution.

[5] A derived word is another example of a (generally) restricted combination whose components can be properly substituted. For instance, nouns of inhabitant in French (Paris+ien /jɛ̃/, New-York+ais /ɛ/, Lill+ois /wa/, Toulous+ain /ɛ̃/) are irregular but can be described as a value of a LF associating city names to the appropriate suffixes. A derivateme is thus a sort of synthetic collocate. The main difference with a true collocation is that the base loses its syntactic properties and acquires new combinatorics.

One last remark about the notion of choice. We do not say that the speaker makes two choices each time (s)he produces a collocation or even a free combination. What we say is that the production of a collocation can be modeled by two consecutive choices: the free choice of the base and the constrained choice of the collocate (while an idiom can never be described as the combination of two choices.)[6]

2.3 Syntaxemes or SSynt units

As mentioned before, I think that the SSynt units are different in nature from DSynt units and can be defined independently. SSynt is only concerned by free combinations and similar ones.

A combination A+B is said to be (*structurally*) *analogous* to a combination A'+B' if and only if A has an equivalent distribution to A', B to B' and A+B to A'+B'.

We call a *syntagm* every combination of linguistic signs that is free or analogous to a free combination.[7] A *syntaxeme* is a homogeneous collection of linguistic signs that are maximal among the signs that are not syntagms. In other words, syntaxemes are the bricks composing the syntagms: every syntagm can be decomposed into a free (or analogous) combination of syntaxemes, while a syntaxeme cannot be decomposed.[8]

A syntaxeme is not *stricto sensu* a linguistic sign (that is, roughly speaking, a correspondence between a meaning and a form), but a collection of similar linguistic signs. First, we put together allomorphs, that is, signs which are in complementary distribution and express exactly the same meaning, like the *-ed* of *cried* and the phonetic alternation that links *ran* to *run*. Second, we put together the signs having the same form(s), related meanings and compatible distributions. Consequently, a syntaxeme is a bundle of signs. Let us take the French verb ALLER 'go' as an example. It has four allomorphs (it is one of the most irregular French verbs): *all+ons* = $\text{ALLER}_{\text{ind, pres, 1pl}}$, *v+ont* = $\text{ALLER}_{\text{ind, pres, 3pl}}$, *i+rons* = $\text{ALLER}_{\text{ind, fut, 1pl}}$, *aill+e* = $\text{ALLER}_{\text{subj, 1sg}}$. And it has many different senses including the base sense 'go' (*J'allais à l'école* 'I went to school'), a meaning of feeling (*Comment allez-vous ?* 'How are you?'), and a use as auxiliary of the future (*Je vais partir* 'I will leave'). The two faces of ALLER are independent: each of its meanings can be expressed by each of its forms and the choice of the form does not depend on the meaning, but on the context (the grammemes combined with the verb).

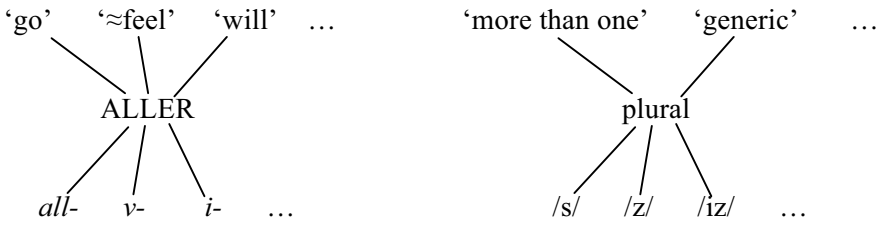


Figure 1. The lexeme ALLER and the grammeme plural as bundles of signs

We consider two types of syntaxemes: lexemes and grammemes.[9] A lexeme is a bundle of lexical units having the same forms and whose meanings are related. This definition of the term, which is adopted by

---

[6] This entails many differences between phrasemes and collocations. In a collocation each LU can be modified independently, allowing various manipulations. Cf for the collocation *make a decision*: *make two decisions, make another decision, the decision I made yesterday.*

[7] The collocation *heavy rain* or the idiom *kick the bucket* are not free combinations, but they are structurally analogous of free combinations like *heavy ball* or *kick the ball*.

[8] In *Mary kicks the ball of Peter*, the syntagms are: *kick-s, ball-∅, the ball, of Peter, the ball of Peter, Mary kicks, kicks the ball, Mary kicks the ball, kicks the ball of Peter,* and *Mary kicks the ball of Peter*. The syntaxemes are: *Mary, kick-, -s, the, ball-, -∅, of, Peter*.

[9] Some derivatemes, like *-er, -able* or *-ness*, combine very freely and could be considered as syntaxemes, but the majority of derivatemes are not syntaxemes.

other authors (see Cruse 1986:49), corresponds to the *vocable* of MTT, which is used as an entry for the ECD dictionary, but not as syntactic unit. Grammemes also are bundles of signs, as noticed in Mel'čuk (1993:vo.1, 278).[10] For instance, the plural in English has several forms (/s/, /z/, /əz/) and several senses ('more than one' as in *I saw raccoons in my garden* or a generic value as in *Whales are mammals*), and each meaning can be expressed by each form.[11]

The fact that the two faces of a syntaxeme are independent means that the correspondence between the syntaxemes and their meanings and the correspondence between the syntaxemes and their forms can be considered independently of each other. There is no need to disambiguate the syntaxemes in SSyntS because knowing what sense of the syntaxeme is considered does not play any role in the correspondence between the SSyntS and the phonological level.[12] Moreover, when the MTM is used in analysis, it is clear that is impossible to disambiguate the syntaxemes without using the lexicon of SUs, that is without activating the Sem-SSynt interface. This is exactly the task of the Sem-SSynt correspondence, when it is used in analysis, to recognize the SUs and to compute the meaning of the (configurations of) syntaxemes.

Another consequence of signs' organization in bundles is that it is sufficient to define the form of SUs in terms of syntaxemes, because their particular phonological realizations only depend on what syntaxemes compose them. In other words, SUs can be defined as subsemiotic entities (Lareau 2008) or half linguistic signs (Kahane 2002): the correspondence between a semanteme and a configuration of syntaxemes forming its signifier. It is what is meant when we call SUs *deep linguistic signs*.

A *phraseme* is an SU whose signifier is composed of several syntaxemes.

We will now look at different cases of SUs that are never considered in the traditional presentations of MTT (Mel'čuk 1974, 1988, to appear, Mel'čuk & Wanner, to appear)

2.4 Grammatical units

A *grammatical unit* [GU] is an SU that is indissociable from a syntactic class of LUs. A GU is freely chosen among a small set of other GUs combining with the same class of LUs. Any choice of a LU in this class imposes to choose one of the GUs commutating with each other. Generally a GU is expressed in the SSyntS by a *grammeme*, that is, by an inflectional syntaxeme.

Our GUs are different in two ways from the deep grammemes that are used in traditional DSyntSs. First a GU is different in nature from a grammeme: it is not a bundle of signs, but a deep linguistic sign with its clearly identified meaning (see Lareau 2008).

Second a GU can be a phraseme (Mel'čuk 1964, 1993:vo.4, Beck 2007). An example is French *conditionnel* (Lareau 2008). From the morphological point of view, the marker of this tense is composed of the morpheme of the future /r/ followed by the morpheme of the past /j/ (written *i*) and in some cases it corresponds to a future in the past as illustrated by (1) (and is then the combination of two GUs):

---

[10] Mel'čuk never exploited the fact that grammemes are organized in bundles to define them. He prefers to consider that the grammemes are "significations", which is a notion not very clear and, I think, less useful than the notion of syntaxeme (see Lareau 2008 for a discussion). One of the reasons Mel'čuk does not want to have a common notion for lexeme and grammeme is that he thinks that grammemes do not have their own combinatorics and that all their "syntactic" properties are attached to their inflectional category (personal communication). We completely disagree with this idea. For instance, the infinitive and the indicative mood in English have very different combinatorics: not only does the infinitive not combine with tense or agreement, but an infinitive verbal form has a completely different distribution than a finite form and we think that this particular distribution is part of the combinatorics of the grammeme. A important part of the grammar is the combinatorics of grammemes (and grammatical lexemes).

[11] More precisely the generic value is expressed by a combination of the plural and a zero determiner.

[12] I really want to insist on that point because it is serious point of disagreement with the traditional presentation of MTT. The major argument to consider with LUs in the SSyntS is that the SSyntS depends on the LU and not just on the syntaxeme expressing it (for instance, the different sense of ALLER have very different subcategorizations: ALLER_go governs a locative preposition group, while ALLER_will governs an infinitive verb). But this is taken into account by the Sem-SSynt interface and, as soon as the SSyntS is build, we do not need to remember what sense of the verb has given this SSyntS for the next calculations.

(1) **a.** *Nous pens+ons que nous viend+r+ons* 'We think that we will come'
b. *Nous pens+i+ons que nous viend+r+i+ons* 'We thought that we would come'

But in its main uses it has an indivisible value, as in *Nous aimerions venir* 'We would like to come' or *Pierre viendrait* (*d'après Marie*) 'Pierre should come (according to Marie)' and it forms a single GU. (Note that *conditionnel* is considered as a single grammatical morpheme by the traditional grammar of French.)

When a grammeme is part of an idiom, like in ⌜CLOSE one's EYE_{pl}⌝ (*He preferred to close his eyes to the possibility of war*) or ⌜CAT_{pl} AND DOG_{pl}⌝ (*It is raining cats and dogs*), it does not appear in the DSyntS. Mass nouns are a limit case where it can be asked whether they form an idiom with a grammeme. They cannot vary in number: most are singular (*water, furniture, anger, news*) and some are plural (*oats,* Fr. *gens* 'people'). In this case the grammeme of number is meaningless and is part of the signifier of the corresponding SU. I think that it must not appear in the DSyntS (contrary to what Mel'čuk & Wanner (to appear) do).

Another example of a phraseological GU is given by the French generic (Kahane 2006b):

(2) **a.** *J'aime le poisson* 'I like fish (the food)'
b. *J'aime les poisons* 'I like fishes (the animals)'

Contrary to English, the generic is expressed in French by using the definite article (LE). More precisely, for countable nouns, the generic is a phraseme combining the definite with the plural: POISSON**a**$_{\text{gener}}$ ⇒ LE_{pl} $\text{POISSON}_{\text{pl}}$ and POISSON**b**$_{\text{gener}}$ ⇒ LE_{sg} $\text{POISSON}_{\text{sg}}$.[13]

2.5 Systematic polysemy links

Systematic polysemy (Barque 2008) gives interesting examples of SUs. Polysemy is said to be systematic when it is not only regular (Apresjan 1973) but when it applies freely to the LUs of a whole semantic field. For instance (countable) nouns denoting animals can be systematically associated to a (mass) noun denoting their flesh. The most common ones are lexicalized (SHEEP → MUTTON, OX/COW → BEEF, FISH**a** → FISH**b**) but any animal noun can be converted into a singular mass noun meaning flesh (*I ate seal once and shark twice*). This conversion is probably a particular case of a more general polysemy link described by Pelletier (1975) and named the *universal grinder*, which maps countable nouns into mass nouns denoting an extract, like in *Peter has egg on his coat*. In the other direction, we have the *universal sorter* (Bunt 1985), which maps a mass noun onto a countable noun meaning 'a type of' (WINE → *a wonderful wine, American wines*), and the *universal packager*, which maps a mass noun onto a countable noun meaning a conventional portion (*two beers, an orange juice*). I think that these conversions, as soon as they combine freely, are SUs and must appear in the DSyntS. Systematic polysemy links are considered as special cases of derivational units and noted D_{X}, where X is the (metalinguistic) name of the operation.

(3) **a.** *I drank wine*: I←I—$\text{DRINK}_{\text{act,past}}$—II→ $\text{WINE}_{\text{indef}}$
b. *I tasted a wonderful wine*: I←I—$\text{TASTE}_{\text{act,past}}$—II→$[\text{D}_{\text{SORTER}}\oplus\text{WINE}]_{\text{sg,indef}}$—MOD→WONDERFUL

Note that the mass noun WINE should not have been inflected in number without being "derived" and that the GU indefinite is expressed by a zero morpheme for mass nouns and by A for countable nouns.

2.6 Government patterns

One of the most challenging questions in the recent literature is the status of the government patterns [GPs] as SUs: To what extent can GPs be considered to freely combine with LUs? There are some uses of lexical items that we do not want to store in the lexicon (Mel'čuk, to appear) like SNEEZE in (4).

(4) *Bob sneezed the napkin off the table*.

---

[13] POISSON**a** and POISSON**b** are SUs: they are two senses of the syntaxeme POISSON. Normally, in DSyntS, we should disambiguate every LU by adding lexicographic numbers, as well as every GU.

For these uses, we consider that the LU corresponding to the standard use has combined with a new government pattern. This GP is noted GP_X where X is the most representative verb with this GP:

(5) BOB←I—[SNEEZE⊕GP_{MOVE}]$_{act,past}$—II→$NAPKIN_{sg,def}$
—III→OFF—II→$TABLE_{sg,def}$

An extremist position, which, I think, underlies Construction Grammar (Goldberg 1995), is to consider that every GP is potentially an SU and that what is commonly considered as a LU in MTT is in fact the combination of a lexical item with a GP. Most of these combinations are not very free and they cannot be considered as two separate choices, but they can be viewed as "phrasemes" combining a lexical item and a GP. For instance, $FAX_{(V)}$ might have been created from a freezing of $FAX_{(N)}$⊕GP_{SEND}. Such a description means that a GP comes with its own meaning and that a big part of the meaning of a LU is in fact borne by its GP. For instance, GP_{MOVE} attributes, to every LU P accepting this GP (like MOVE, PUT, PUSH, BREAK, SNEEZE …) , the meaning: 'X P-s Y in Z' = 'acting on Y by P-ing || X causes Y to be moved in Z'. GP_{PUT} is a sort of hypernym of all the LUs having this GP.

2.7 Constructional units

We call *constructional units* [CUs] the SUs whose signifier cannot easily be attributed to lexemes or grammemes, but must be rather attributed to a whole syntactic construction. I prefer this term to the more general term *construction* also used for designating GPs. Every sentence is headed by a CU corresponding to the speaker's choice to utter a declarative sentence rather than an interrogative or an exclamative one.[14] These SUs are dealt with in a particular way in traditional MTT: They have a meaning (something like 'the speaker asks the hearer S' for an interrogative sentence S), but they correspond in the DSyntR to a prosodic unit which is put in a special field and is not directly related the DSyntS. I do not see any good reason to do that; even if they are mostly expressed by the prosody, they also control special word orders and particular markers and they can intervene in several points of the DSyntP (Fig. 2a,b).[15]

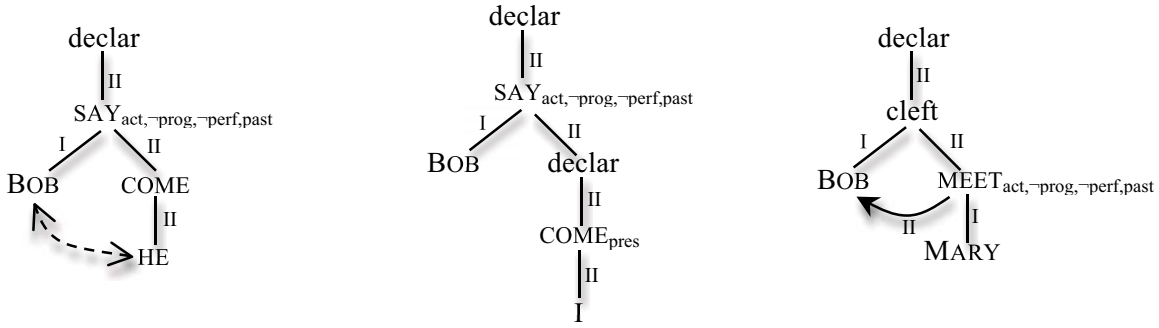
Another very important set of CUs concerns the communicative structure. For instance English has a dedicated construction for expressing a prominent rheme: clefting. There are many arguments for considering this construction as an SU. Its signifier is quite rich, combining many grammatical syntaxemes and acting on the whole SSyntS of the sentence. From the semantic point of view, it not only modifies the information packaging, but it also modifies the presuppositions in a manner comparable with some GUs. Cf. the following parallel between definiteness and clefting in French: *C'est Bob qui vient* 'It is Bob who is coming' ≅ 'The one who is coming is Bob' vs *Il y a Bob qui vient* ≅ 'Bob is coming' in the sense 'Someone who is coming is Bob'. A last argument for considering the clefting as a node in the DSyntS (Fig. 2c): The second actant of the clefting is optional (*It is Bob*). It would be very difficult to give a reasonable DSyntS of a sentence like *I think it is Bob* (answering *Who is coming?*) without having introduced the "cleft" node.

---

[14] Maybe these SUs must be considered as GUs: their choice is imposed as soon as we utter a new sentence and they are chosen among a finite set of alternatives. But they are less clearly associated to a particular class of LUs.

[15] Some remarks about the DSyntSs of Fig. 2. There is no past tense on the embedded verb in Fig. 2a because it is not an SU: the speaker did not choose this tense, it was imposed by an agreement rule. But there are other GUs expressed by tense grammemes, like the habitual, that must appear in this position: *Bob said that he comes* ($COME_{HAB}$).

In the DSyntS of Fig. 2b we choose to leave the pronoun HE, which is coreferential with BOB. In the traditional DSyntS the pronoun would have been replaced by a copy of BOB. This is not a possible solution with our definition of the DSyntS. Only two solutions are possible. First solution: *he* is an SU's expression and it must appear as a node in the DSyntS. This SU is a very special SU, functioning like a substitute, clearly different from the SU BOB, even if in this context it corresponds to the same Sem node. Second solution: *he* does not result from a choice of the speaker, but it is imposed by a rule of pronominalization. In this case we should consider that there is no specific SU corresponding to *he* and that *he* and *Bob* correspond to the same SU; see Kahane (2003b) where we defend this position which makes the DSyntS a dag rather than a tree. In any case a deeper study of pronominalization is needed, which to my knowledge has never been rigorously modeled in MTT.



a. *Bob said that he came.* **b.** *Bob said: "I come."* **c.** *It is Bob who Mary met.*

Figure 2. DSyntSs.

3 The DSynt Structure

We define the *DSyntS* as the structure indicating how the SUs combine together when an utterance is produced.

In many respects, this gives us a structure very similar to the traditional DSyntS (Mel'čuk 1988, to appear) and it is why I think the structure we defined is what the traditional DSyntS tries to capture, even if it is defined in another way. But above all, if I define the DSyntS like this, it is because it is what we want to model: How a speaker chooses and combines SUs in an utterance. (We will come back to this central question in Section 4).

In the previous section we saw some differences between the traditional DSynt units and the SUs we introduced here, but the core remains the same. In this section we focus on the relations between these units in the DSyntS.

3.1 Combination of LUs with GUs

Grammatical units are signifying units in the same way as lexical units and they are true elements of the DSyntS. They are traditionally presented as subscripts of LUs. This representation does not make explicit that the GUs are objects of the DSyntS in the same manner as LUs and it does not clearly indicate how the GUs combine together or with the LUs.[16]

A verbal form like *will have been being eaten* is traditionally described at the DSynt level a $\text{EAT}_{\text{pass, prog, perf, fut}}$.[17] How must this notation be interpreted? I think that the main motivation of this notation, giving a different status to LUs and GUs, is to show that a LU and the GUs that combine with it make a whole and that it is this group that combines with other groups of the same type. I thus want to emphasize that this means that the structure under analysis is not really a tree, but has nuclei in the sense of Tesnière (1959) (see Kahane 1997 for a formalization with bubble trees).[18] I interpret the traditional notation as shown in Fig. 4, where the SUs constituting the verbal form combine in a strict order and where the whole combination forms a bubble, which is a node of the DSynS linked to other nodes by DSynt relations (see next

[16] In this presentation we will not get into the formalism and give formal rules. But when we say that the GUs are objects we think of the modeling of the Sem-SSynt interface we have done in Meaning-Text Unification Grammar (Kahane 2002, 2006, Kahane & Lareau 2005, Lareau 2008) where objects are elements really handled by the grammar.

[17] It is not exactly what Mel'čuk (to appear) proposes. We consider that the indicative mood, although it is a grammeme, is not a GU, but is part of the declarative or interrogative construction for a main verb or part of the government for a subordinated one. We are not sure that the perfect and the progressive are grammemes, but we will not discuss this point here (see Lareau 2008 for a related discussion about the *passé composé* of French).

[18] I am not totally convinced that this complication of the structure is needed and I never saw arguments specifically arguing that.

section). Arguments for considering such a strict order in the combinations come both from the form and the meaning. Concerning the form, it is clear that these GUs form an SSynt chain of verbs in this order.[19] Concerning the meaning, *will have been being eaten* does not mean that the eating process will occur in the future, but that it will be completed when it will be considered in this future and so 'future' combines with 'perfective' and not with 'eat'. In the same way, it is not the eating which will be completed but the being eaten and 'perfective' combines with 'progressive' and not directly with 'eat'.

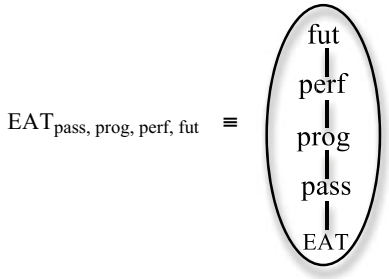


Figure 4. Interpretation of the traditional encoding of GU's relations

3.2 DSynt relations

The DSyntS indicates how the SUs of an utterance have combined. A DSynt relation is a link in the DSyntS between two SUs showing that these SUs have combined and how they have combined. For instance, if we postulate that GUs combine with others SU in a special way, this justifies the use of a special notation for this link (see the previous section). We focus now on the links between LUs, or more exactly between the nuclei formed by a LU and its "satellites" (GUs, but also derivatemes or GPs).

From a purely formal point of view, DSynt relations do not need to encode much information; we just need to distinguish, between the various combinatorial possibilities a LU has, which combination has been used.[20] Any additional information is only useful for the sake of readability or for capturing some universal regularities in the shape of DSyntSs.

When SUs combine, they combine both at the Sem and at SSynt levels. From the Sem viewpoint, we must indicate, which SU is the Sem predicate and which one is the Sem argument. From the SSynt viewpoint, the combination is also asymmetric, and we must indicate which SU is the SSynt governor and which one is the SSynt dependent. These give us four possibilities (Fig. 5). We put the syntactic governor on the top of the link as usual. When there is no SSynt relation between the nodes we use a curved arrow from the predicate to the argument.

Sem and SSynt (same direction)	Sem and SSynt (opposite direction)	only Sem	only SSynt
			

Figure 5. The four types of DSynt relations between LUs.

[19] The fact the GUs combine in a strict order is evident here, because they are expressed by separate words forming a chain of SSynt dependencies. This was first stated by Pollock (1989), who proposes that grammemes head different projections of the verbal form embedded in each other.

[20] For instance, in Tree Adjoining Grammar [TAG], SUs are modeled by elementary phrase structure trees, which can combine by two operations, substitution and adjoining. A link in a TAG derivation structure indicates which operation has been used, which tree is the host (combinations are asymmetric) and the address of the host's node where the other tree substitutes or adjoins (Vijay-Shanker 1987).

An SU syntactically depending on one of its Sem argument is called a *modifier*; a modifier is placed below its governor; the DSynt-link is labeled MOD (rather the too exotic traditional ATTR). SUs which are both Sem arguments and SSynt dependents of a given SU are called its *actants* (Tesnière 1959). An SU can have several actants and therefore we need to distinguish the different actancial DSynt relations. From a formal viewpoint, the way we do that does not really matter; the numbering following the syntactic saliency which is traditionally adopted is ok.[21]

Contrary to the traditional DSyntS which favors the SSynt dependencies in case of mismatches with the Sem, I think that both SSynt and Sem dependencies must be kept in the DSyntS. Raising verbs give a typical example: in *Bob seems to sleep*, BOB is the SSynt subject of SEEM, while 'Bob' is a Sem argument of 'sleep' but not of 'seem' (Fig. 6, left). The label +I indicates that this relation is only SSynt, its Sem counterpart being realized elsewhere (here between SLEEP and BOB).

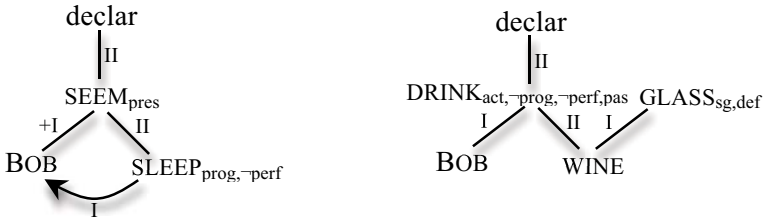


Figure 6. DSyntSs of *Bob seems to sleep* and *Bob drank a glass of wine.*

Another problem is illustrated by predicative adjunction as in *Bob drank a glass of wine*. Here WINE is the DSynt actant of GLASS (it is both its SSynt dependent and its Sem argument). Nevertheless, the Sem argument of 'drink' is clearly 'wine' (Bob drank wine and not glass). Therefore we consider that DRINK essentially combines with WINE and the special behavior of GLASS allows its syntaxeme to intervene between DRINK and WINE at the SSynt level (Fig. 6, right). The related very interesting problem posed by extraction will not be studied here (Kahane & Mel'čuk 1999, Kahane 2002).

4 Status of the DSyntS

4.1 Monolingual model

Traditionally the DSynt level is presented as an intermediate level between the Semantic and Surface Syntactic level. The synthesis of a sentence begins with a Semantic Representation [SemR] and the Sem-DSynt interface (also called the Sem-module) builds corresponding DSyntRs, which in turn correspond to SSyntRs via the DSynt-SSynt interface (or DSynt module). In such presentations it is never said where SemRs come from. I do not think that semantemes are chosen before, and independently of, SUs whose signifiers they are. I think that, given some communicative goals and considering some referents in the world, the utterance is built directly by choosing SUs whose signifiers cover the situation we want to describe or the communicative goals we want to reach. In other words, a SemR cannot be envisioned without having produced a corresponding utterance, even if the speaker does not pronounce it aloud. A speaker does not select semantemes, but instead directly selects SUs (that is, deep linguistic signs), whose combination produces simultaneously a SemR and a SSyntR.

What is proposed does not contradict the traditional view of the architecture of a traditional Meaning-Text model [MTM]. The different levels of representation of an MTM are still here, in the same order.

[21] In fact, Mel'čuk numbering is more complicated: The most salient actant of a verbal SU is numbered II if it is not a SSynt subject and the agent complement of a passive is numbered II despite the fact it is generally less salient than the actant III. If the first adjustment allows a syntactically more homogenous set of actants I, the second adjustment has the opposite result. The conventions proposed by Mel'čuk have the ambition that each DSynt actantial relation covers a universally homogeneous set of SSynt relations, but the conventions adopted by Bresnan (2001) for the functional structure of LFG are certainly better from this viewpoint (and much more widespread).

But we do not consider the DSyntR functions as an intermediate level. I think that there is a direct correspondence between Sem and SSynt and that these two levels are put in correspondence when the SUs are chosen by the speaker.

From a practical point of view, there is no gain in considering two separate modules (I mean the Sem-DSynt and DSynt-SSynt interfaces). In the traditional presentation of MTT, there is only one lexical resource for these two modules—the Explanatory Combinatorial Dictionary—and it would not make sense to break it into two separate dictionaries, one for each module. Entries of an ECD are LUs and each entry describes the SSynt realization of each Sem argument, that is the correspondence between the Sem representation of a LU and its SSynt representation. If the two correspondences (from Sem to DSynt and then from DSynt to SSynt) were established independently, each lexical entry would be called twice; this is the case for instance in the text generation system described in Lareau & Wanner 2007. It seems more reasonable to have only one module where the DSyntS is built as a witness of the way we go from the Sem to the SSynt level, calling each lexical entry only once.

Kahane (2003a) formulated this in other terms: DSyntSs are derivation structures of the Sem-SSynt interface. A *derivation structure* is a structure recording how the rules of a grammar have combined to derive a sentence; the nodes of the structures are labeled by rules and the links between the nodes indicate how the rules have combined. The notion of a derivation tree, coming from CFGs (Chomsky 1957), has been extended to TAG by Vijay-Shanker (1987) (see note 20). Rambow & Joshi (1992) showed that TAG derivation trees are similar to DSyntSs of MTT.[22] SU entries can be assimilated to correspondence rules between the Sem and SSynt levels (the correspondence between the two faces of the sign) and the DSyntS, which is by definition the structure recording how the SUs combine, can thus be viewed as the derivation structure of Sem-SSynt correspondence. MTT has been formalized in this way under the name of Meaning-Text Unification Grammar (Kahane 2001, 2002, Kahane & Lareau 2005, Lareau 2008), inspired by TAG and HPSG (see Kahane 2009 for an HPSG-wise dependency grammar).

4.2 DSyntSs from the viewpoint of paraphrasing and translation

Paraphrasing is expressing (nearly) the same meaning by different means. And translation can be view as a particular case of paraphrasing, where different means are different languages. Most of paraphrasing and translation can be achieved by replacing SUs with other SUs (Mel'čuk 1988, Milićević 2007, Kahane 2007) and this is why the DSyntS is so central in paraphrasing and translation. And it is exactly because the DSyntS is not an intermediate level but a derivation structure that it is the structure used in paraphrasing rules. Because a paraphrasing rule is a rule explaining how to differently realize the correspondence between a given meaning and its linguistic realizations. To paraphrase is to change the derivation.

The consideration of paraphrasing and translation gives a particular view on the DSyntS and raises an interesting question: what is relevant to encode in the DSyntS from the viewpoint of paraphrasing and translation? It is probable that many choices concerning the DSyntS have been made considering the paraphrasing, notably because MTT grew in the context of Machine Translation. This is why Mel'čuk argued that the DSynt relations must be universal, which is not very justifiable from the viewpoint of a monolingual linguistic model and not necessarily useful for modeling paraphrasing or translation.[23] The choice of relegating the grammatical elements as subscripts to LU names and considering them as grammemes rather than true SUs is also a result of the desire to mask idiosyncrasies of each language and to concentrate on lexical translation.

---

[22] As seen in Section 3.2, the way SUs combine together does not necessarily yield a tree. The traditional DSyntS favors the SSynt dependencies, while the TAG rather favors the Sem dependencies (Candito & Kahane 1998).

[23] In a paraphrasing rule replacing an SU X by another SU X' (it can involve more SUs than that but this does not change the problem) we need to explain how the SUs potentially linked to X will be stuck back on X'. The rule must tell which actant of X corresponds to which actant of X', but this can be done no matter how we name them, even for LFs.

5 Conclusion

I think that the DSyntS is a crucial linguistic structure that should have a central role in any attempt to model natural languages. Although it appeared more than forty years ago, it remains sufficiently unknown outside the MTT community. One of the reasons for this is the way the DSyntS is presented in MTT publications, as a level of representation internal to the theory and not characterized on its own. Moreover, even if the general principle of the DSyntS is easy to defend, most of the particular choices of representation are not really discussed. In fact it is very difficult to argue for a particular choice of representation if we have not defined what the DSyntS is about and what it encodes. In this paper I have tried to advance in this direction, proposing that the DSyntS is the structure encoding how the signifying units combine with each other when a sentence is uttered. The structure so defined is not very far from the traditional DSyntS. The few differences we obtain point to some aspects of the DSyntS that should be discussed further. The discussion is open and many problems remain to be solved.

Acknowledgements

This paper has benefited from helpful comments of François Lareau, Igor Mel'čuk, Owen Rambow, and two reviewers. I am particularly grateful to Bob Coyne who helps me for the English version.

References

Apresjan, Jurij. 1973. Regular polysemy, *Linguistics 12*, 5-32.

Barque, Lucie. 2008. *Description et formalisation de la polysémie régulière*. PhD thesis, Université Paris 7.

Beck, David. 2007. Morphological phrasemes in Totonacan inflection, *Proceedings of MTT*, Klagenfurt.

Bresnan, Joan. 2001. *Lexical-Functional Syntax*, Blackwell, Oxford.

Bunt, Harry C. 1985. *Mass Terms and Model-Theoretic Semantics*. Number 42 in Cambridge studies in linguistics. Cambridge University Press.

Candito, Marie-Hélène, & Sylvain Kahane. 1998. Can the derivation tree represent a semantic graph? An answer in the light of Meaning-Text Theory, *Proceedings of TAG+4*, Philadelphia, 21-24.

Chomsky, Noam. 1957. *Syntactic Structures*, MIT Press, Cambridge.

Cruse, D. A. 1986. *Lexical Semantics*, Cambridge University Press, Cambridge.

Ducrot, Oswald. 1995. Les unités signifyings [The signifying units]. In Ducrot O. & J.-M. Schaeffer, *Nouveau dictionnaire encyclopédique des sciences du langage*, Seuil, Paris, 358-365.

Goldberg, Adele. 1995. *A Construction Grammar approach to argument structure*, University of Chicago Press, Chicago.

Kahane, Sylvain. 1997. Bubble trees and syntactic representations, in Becker & Krieger (eds), *Proc. 5th Meeting of the Mathematics of Language (MOL5)*, DFKI, Saarbrücken, 70-76.

Kahane, Sylvain. 2002. *Grammaire d'Unification Sens-Texte : vers un modèle mathématique articulé de la langue*, Habilitation thesis, Université Paris 7.

Kahane, Sylvain. 2003a. On the status of the Deep Syntactic Structure, *Proceedings of MTT*, Paris.

Kahane, Sylvain. 2003b. The Meaning-Text Theory, *Dependency and Valency, Handbooks of Linguistics and Communication Sciences 25 : 1-2*, De Gruyter, Berlin/NY, 32 p.

Kahane, Sylvain. 2006. La distribution des articles. In M. Charolles, N. Fournier, C. Fuchs & F. Lefeuvre (eds.), *Parcours de la phrase - Mélanges offerts à Pierre Le Goffic*, Ophrys, Paris, 159-174.

Kahane, Sylvain. 2007. A formalism for machine translation in MTT, including syntactic restructurings, *Proceedings of MTT*, Klagenfurt.

Kahane, Sylvain. 2009. On the Status of Phrases in Head-driven Phrase Structure Grammar - Illustration by a Totally Lexical Treatment of Extraction, in Alain Polguère & Igor Mel'čuk (eds), *Dependency in Linguistic Description*, Language Companion Series, 111-150, John Benjamins, Amsterdam/Philadelphie.

Kahane, Sylvain, & François Lareau. 2005. Meaning-Text Unification Grammar: modularity and polarity, *Proceedings of MTT*, Moscow, 23-32.

Kahane, Sylvain, & Igor Mel'čuk. 1999. La synthèse sémantique ou la correspondance entre graphes sémantiques et arbres syntaxiques – Le cas des phrases à extraction en français contemporain*, T.A.L.*, 40:2, 25-85.

Kahane, Sylvain, & Alain Polguère. 2001. Formal Foundations of Lexical Functions, Proceedings of the *Workshop on Collocation, ACL 2001*, Toulouse, 8 p.

Lareau, François. 2008. *Vers une grammaire d'unification Sens-Texte du français : le temps verbal dans l'interface sémantique-syntaxe*, PhD thesis, Université de Montréal & Université Paris 7.

Lareau, François, & Leo Wanner. 2007. Towards a Generic Multilingual Dependency Grammar for Text Generation, *Proceedings of the GEAF07 Workshop*, Standford.

Martinet, André. 1960. *Éléments de linguistique générale*, Paris.

Mel'čuk, Igor. 1964. Obobščenie ponjatija frazeologizma (morfologičeskie "frazeologizmy") [A generalization of the concept of phraseoligism (morphological "praseologisms")]. In L.I.Rojzenzon (ed.), Materialy konferencii *"Aktual´nye voprosy sovremennogo jazykoznanija i lingvistič]eskoe nasledie E.D. Polivanova"*, vol. I, Samarkand : SamGU, 89-90.

Mel'čuk, Igor. 1988. *Dependency Syntax: Theory and Practice*, SUNY Press, Albany.

Mel'čuk, Igor. 1993. *Cours de morphologie générale*, CNRS, Paris/Presses de l'Université de Montréal.

Mel'čuk, Igor. 2001. *Communicative Organization in Natural Language: The Semantic-Communicative Structure of Sentences*, John Benjamins, Amsterdam/Philadelphia.

Mel'čuk, Igor. To appear. *Semantics: From Meaning to Text*, vol. 1 - 3, John Benjamins, Amsterdam/Philadelphia.

Mel'čuk, Igor, & Leo Wanner. 2006. Syntactic Mismatches in Machine Translation. *Machine Translation*. 20(2):81-138.

Mel'čuk, Igor, & Leo Wanner. To appear. Morphological Mismatches in Machine Translation. *Machine Translation*. 59 p.

Milićević, Jasmina. 2003. La paraphrase - Modélisation de la paraphrase langagière, Peter Lang.

Pelletier, Francis J. 1979. *Mass Terms: Some Philosophical Problems*. Reidel, Dordrecht.

Pollock, Jean-Yves. 1989. Verb Movement, Universal Grammar, and the Structure of IP, *Linguistic Inquiry*. 20:365-424.

Rambow, Owen, & Aravind Joshi. 1992. A Formal Look at Dependency Grammars and Phrase-Structure Grammars, with Special Consideration of Word-Order Phenomena, *International Workshop on The Meaning-Text Theory*. Darmstadt. Arbeitspapiere der GMD 671.

Saussure, Ferdinand de. 1916. *Cours de linguistique générale*, Paris.

Tesnière, Lucien. 1959. *Eléments de syntaxe structurale*, Klincksieck, Paris.

Vijay-Shanker, K.1987. *A Study of Tree Adjoining Grammars*, PhD thesis, University of Pennsylvania.

Le rôle du verbe auxiliaire dans l'alternance de codes kisongye/français

Sébastien Kitengye Sokoni

Doctorant à l'Université de Kinshasa (R.D.C.)

samsoki@yahoo.fr

Résumé – Abstract

L'objectif de cet article est de montrer, par l'étude du rôle du verbe auxiliaire dans l'alternance kisongye/français, que le Modèle Sens-Texte est aussi applicable à l'alternance de codes. Ainsi, nous démontrons dans l'interface sémantique-syntaxe que le recours à la structure à verbe auxiliaire est un cas typique de l'alternance de codes qui relativise la théorie de la langue matrice comme établie par Myers-Scotton (1993 a&b).

This article attempt to apply MTT to codeswitching by determining how important is auxiliary verb in kisongye/french codeswitching,. Thus, in Syntax-Semantic interface we demonstrate that auxiliary verb structure is specific to codeswitching which minimize matrix language theory as established by Myers-Scotton (1993 a&b).

Mots clés – Keywords

Théorie Sens-Texte, Alternance de codes, Verbe auxiliaire, Sémantèmes mixtes, Langue matrice, langue enchâssée.

Meaning-Text Theory, Codeswitching, Auxiliary verb, Mixt semantemes, Matrix Language, Embedded language.

1 Introduction.

Etudier le processus de synthèse de l'alternance de codes en simulant la capacité de l'être humain à exprimer un contenu dans un énoncé à code mixte par le recours au verbe auxiliaire, tel est le but poursuivi dans cet article.

Dans une telle synthèse, il ne s'agit pas de s'interroger a priori sur le rapport entre langues en présence comme c'est le cas dans le processus de génération multilingue. Il s'agit de passer par la description des structures à codes alternés manipulant les verbes auxiliaires (moyen) pour déterminer les rapports entre langues (fin). Bref, il s'agit de fixer ce rapport a posteriori.

Dans la mesure où le processus de synthèse se réalise simultanément dans les deux langues pour un énoncé donné, la dépendance entre celles-ci est possible. L'une devrait dépendre de l'autre du point de vue syntaxique. Il est possible d'observer une co-dépendance qui pourrait engendrer des structures ne relevant pas ou partiellement de l'une ou l'autre langue en présence.

La question que l'on se pose est celle de savoir si les énoncés à verbe auxiliaire sont produits en utilisant uniquement les règles des deux langues ou celles de la résultante de leur contact. L'analyse du rôle des verbes auxiliaires de l'alternance de codes kisongye/français par l'approche Sens-Texte consistera à

montrer qu'il existe des constructions propres à la résultante du contact des deux langues qui n'appartiennent à aucune des deux langues qui ont servi à les construire. Cela permet de vérifier la validité des théories éprouvées comme celle de la langue matrice[1] (Myers-Scotton, 1993b).

Pour ce faire, il s'agira d'écrire des représentations syntaxiques d'énoncés mélangeant les deux langues afin de déterminer celle dont les règles agissent au niveau des structures syntaxiques lorsque l'alternance est construite sur base d'un verbe auxiliaire.

Dans cet article, nous entendons par alternance de codes « la stratégie de communication utilisée par les bilingues. Cette stratégie consiste à faire alterner des unités de longueur variable de deux ou plusieurs codes à l'intérieur d'une même interaction verbale » (Hamers et Blanc, 1983). Nous savons néanmoins que Myers-Scotton (1993b) considère l'alternance de codes comme « sélection par le bilingue (…) des formes provenant d'une langue source (enchâssée) dans la langue matrice au cours du même énoncé ».

Notons que les éléments du corpus présentés dans cette étude proviennent d'un important corpus de 25 heures d'interview enregistré en milieu scolaire songye (à Tshofa) sur un échantillon randomisé de 100 témoins, tous usagers de l'alternance de codes kisongye/français

Dans cet article, nous posons d'abord les fondements de la théorie Sens-Texte appliquée à l'alternance de code avant de montrer le rôle du verbe auxiliaire dans une alternance particulière.

2. Quelques aspects de la grammaire du kisongye.

Dans cette section, nous présentons sommairement quelques aspects de la grammaire du kisongye utile à cette étude. Il s'agit de la morphologie du verbe et de sa dérivation.

De prime abord, l'on doit noter que la structure générale d'une forme verbale du kisongye est la suivante : PV – Rad – F (Préfixe verbale – Radical – Final) :

(1) Kasongo ádidi : á – dil – i
(Kasongo 3SG pleure PRESENT)
'Kasongo pleure'

Les préfixes verbaux sont attestés dans les formes verbales conjuguées. Ils se répartissent en préfixes des participants (1ère et 2ème personnes du singulier et du pluriel) : nadidi : na – dil – i (je pleure) ; odidi : o – dil – i (tu pleures) ; atudídí : atu – dil – i (nous pleurons) ; anudídí : anu – dil – i (vous pleurez) et en préfixes des classes (classes 1 à 18) : muci úbápónó : u – ba – pon – o (CL3 – PRESENT – tomber) 'l'arbre est tombé'.

L'infinitif du verbe kisongye reçoit le préfixe nomino-verbal ku- de classe 15: kú – sepa : (rire) tandis que l'impératif est une forme sans préfixe : - dila ! (pleure !). La structure du subjonctif est la suivante : PV – Rad. – e : Kasongo adile : a – dil – e (Que Kasongo pleure). C'est la suffixe « – e » qui en est la marque. Les suffixes – ilé et –ang – sont les marques temporelles exprimant le passé : Kasongo básépele : ba – sep – ilé (Kasongo avait ri) ; Kasongo ásepangá : a – sep – ang – a (Kasongo riait). Il faut noter que la finale verbale caractérise aussi la forme verbale du point de vue temporel : Kasongo ákasepa : a – ka – sep – a (Kasongo rira)

[2] Myers-Scotton pose les principes contraignants suivants : le principe du morphème fonctionnel selon lequel ces morphèmes doivent appartenir à la langue matrice ; le principe de l'ordre des morphèmes qui soumet cet ordre au diktat de la langue matrice ; et enfin le principe de blocage qui bloque certaines unités du discours de la langue enchâssée au profit de celles de la langue matrice.

S'agissant de la dérivation verbale, elle se réalise en kisongye au moyen des suffixes : - il – (pour la dérivation applicative) ; - ish – (pour la dérivation causative) ; - an – (pour la dérivation réciproque ; – ik – (pour la dérivation causative) … : kúdila : ku – dil – a (pleurer) ; kúdidila : ku – dil – il – a (pleurer pour… : applicatif) ; kúdidisha : ku – dil – ish – a (faire pleurer : causatif) ; kwidileena : ku – i – dil – an – a (se pleurer : réciproque) -

3 Fondements de la TST appliquée à l'alternance de codes.

Nous partons des fondements de la théorie Sens-Texte monolinguistique pour l'étendre à une situation d'alternance de codes. Nous estimons, à la suite de Mondada (2007), que parmi les défis que pose l'alternance de codes au modèle linguistique, le plus important est que l'alternance amène le problème de la prise en compte non seulement de plusieurs variétés mais encore de plusieurs langues au sein du même énoncé. Pourtant, les modèles de la grammaire sont plus ou moins tacitement basés sur une seule langue considérée comme un système homogène.

Au lieu d'envisager isolément « les morphèmes fonctionnels », « l'ordre des morphèmes » et le « blocage » de certaines unités du discours mixte dans l'alternance de codes comme indices suffisants du rôle de l'une des langues appelée « matrice », nous observons les structures ou plutôt les relations syntaxiques globalement et prenons en considération les différents modules de la TST qui manipulent ces relations.

La description de l'alternance de codes passe par l'étude de la phrase, la proposition, le syntagme et le mot à en considérant que la phrase et le mot sont les unités de base (Mel'čuk, 1997). Il s'agira d'étudier les relations syntaxiques avec en toile de fonds les notions de dépendance et de fonction syntaxique.

Lors de la mise en commun des unités des deux langues, pour faciliter la production des discours bilingues ou plurilingues, le niveau de Représentation Sémantique est unique (RSém). Il reçoit des sémantèmes des langues en présence chez l'individu.

C'est cette RSém qui est considérée comme point de départ du processus Sens-Texte dans l'alternance de codes dans la mesure où, d'après Blanchet (2004), « les plurilingues ne sont pas des pluri-monolingues. Il se compose un seul répertoire linguistique fait d'éléments ailleurs identifiés comme provenant des langues distinctes ».

Comme on le voit, des sémantèmes de chacune des deux langues sont présents dans ces représentations sémantiques. Ce qui provoque l'alternance de codes. Lüdi (1995) écrira : « [C'est de cette façon qu'un locuteur bilingue s'y prend] pour formater plus ou moins simultanément une représentation dans deux ou plusieurs langues »

4 Construction des Représentations sémantique, syntaxique et morphologique en alternance de codes.

4.1 La RSém.

Dans cette construction, la RSém à sémantèmes mixtes incarne la SSém en alternance de codes. C'est à partir d'elle donc que s'opère le choix entre les segments de l'une et l'autre langue. On obtient, pour un énoncé à double argument par exemple, la RSém suivante : 'X_k' ←1— 'Y_f'—2→ 'Z_k' | $\_k$: Langue kisongye ; $\_f$: Langue française ; X, Z : arguments du verbe Y.

Le locuteur construit une Rsém bien formée en mélangeant des sémantèmes des deux langues (cf. supra). Il choisit pour chaque concept un seul sémantème dans une des deux langues. Chaque représentation à

sémantèmes mixtes ou « paraphrase sémantique » correspond à une paraphrase au niveau syntaxique profond et de surface.

4.2 RSyntP, RSyntS et RMorph.

Les RSyntP, RSyntS et RMorph**.** restent identiques à celles du MST général telles qu'elles sont définies par Polguère (1998), Kahane et Lareau (2005). Elles s'appliquent bien à l'alternance de codes avec comme spécificité que le calcul des relations syntaxiques de surface se fait après le choix des segments de l'une ou l'autre langue (lexicalisation) sans modifier les arcs ou les relations syntaxiques. Celles-ci, on le sait, relèvent des dépendances syntaxiques profondes. L'introduction des lexies vides se fait conformément aux contraintes qui résultent des fluctuations entre langues en présence. Cela nous permet de vérifier l'hypothèse d'une structure de la langue matrice qui dicterait ces contraintes à l'alternance de codes. D'après cette hypothèse, dans le constituant LM (langue matrice) + LE (langue enchâssée), tout morphème fonctionnel en relation grammaticale avec le noyau (constituant de base) proviendra de la langue matrice.

La construction de la représentation morphologique profonde dans l'alternance de codes n'est pas non plus spécifique à celle-ci. Elle suit les principes généraux qui sont supposés connus (Gerdes & Kahane, 2004). La tâche principale à effectuer lors du passage RSyntS $\Rightarrow$ RMorphP est la linéarisation de l'arbre syntaxique. Il y a lieu de noter également que dans la RMorphP, la chaîne des lexies est marquée morphologiquement par le calcul de différents accords morphologiques (par exemple accord du verbe avec son sujet grammatical) (cf. Marie Hélène Candito et Sylvain Kahane, 1998). Il y a aussi le calcul de la prosodie syntaxique (de la phrase). La chaîne des lexies de la phrase ne porte que les indications morphologiques pertinentes du faite qu'elles appartiennent aux deux langues. C'est le cas de la mention du nombre. Le genre ne joue un rôle dans ce cadre que lorsque les deux langues en présence y recourent. Il en va de même des indications de nombre et de personne.

5 Le rôle du verbe auxiliaire dans l'alternance de codes kisongye/français.

L'alternance de codes kisongye/français résulte du contact entre le kisongye (langue bantu parlée en République Démocratique du Congo dans la région située entre 23° et 27° de l'Est à l'Ouest et entre 4° et 7° du nord au sud ; elle est codifiée L23 dans le répertoire de Malcom Guthrie (1970) et le français.

À en croire Joshi (1987), le contact de deux langues précitées, langues typologiquement différentes, devrait engendrer une alternance asymétrique avec dominance du kisongye. En effet, dans une telle alternance, la langue matrice est celle dont les morphèmes sont numériquement supérieurs par rapport à ceux de la langue enchâssée. Ainsi, dans l'énoncé (2) ci-dessous, le kisongye est-il la langue matrice parce qu'il comporte quatre morphèmes alors que le français en compte trois :

(2) Maître **bá**rencontrer **balongi boosó mú** champ : ba - rencontrer
Maître 3SG rencontrer CL2élèves tous dans champ.
('Le maître a rencontré tous les élèves au champ'.)

Nous pensons pour notre part que seule la langue dont les relations syntaxiques s'appliquent à l'alternance peut être considérée comme langue matrice. Bien que l'énoncé (2) ci-dessous soit numériquement dominé par le français, la relation de focalisation pseudo-clivée utilisée est celle du kisongye (ici la nasale syllabique N de focalisation). Le locuteur produisant cet énoncé sait qu'il parle kisongye et l'auditeur reconnaît cette langue malgré la présence de trois segments étrangers (du français) :

(3) Ba**manger riz** m**père**. Ba – manger ; N - père
3SG manger riz FOC.- père.
('Celui qui a mangé du riz c'est le papa')

D'où la nécessité de recourir aux relations syntaxiques plutôt qu'au nombre de segments pour déterminer la langue matrice. Comment se comportent la structure à verbe auxiliaire dans une telle alternance ?

Dans la partie suivante, ces verbes sont étudiés dans les structures causative, impositive, réciproque et dans les formes verbales de l'indicatif passé, de l'infinitif habituel, de l'impératif et du subjonctif. Le choix de ces structures est dicté par le fait qu'elles sont obtenues en kisongye par la suffixation non employée en alternance de codes. Il s'agit donc de voir comment l'alternance de codes comble l'impossibilité de suffixer les $Vinf_f$ en kisongye.

5.1 Le verbe auxiliaire dans une structure causative.

La structure causative, formalisée en Rsém : 'x' ←1—'causer—2→'P'—1→'y' (où P est le verbe auquel s'applique la construction causative) est pris au sens que lui donnent Kahane et Mel'čuk (2006).. Cette structure se réalise en kisongye par un biais dérivationnel (recours au suffixe dérivatif « - ish -» :

(4) Natumíkíshá bálongi : °na – tumik – ish – a
1SG travailler-caus. PRES. les élèves.
('Je fais travailler les élèves')

L'alternance kisongye/français recourt au verbe « faire » du français ou à son équivalent « kúkitá » du kisongye comme verbe auxiliaire :

(5) a. Nafaire travailler baélèves.
$1SG_k$ faire travailler $CL2_k$ élèves
Je faire travailler les élèves
('Je fais travailler les élèves')

b. Nakicíshá kútravailler baélèves : Na – kit – ísh - á
$1SG_k$ faire-CAUS. PRES. travailler $CL2_k$ élèves.
Je faire travailler les élèves
('Je fais travailler les élèves')

L'énoncé (5b) ci-dessus se prête à la correspondance suivante : 'kukita'— 2→ 'Y' ⇔ KUKITA — objet→ KU + Y. Y est le verbe du français (V_f) tandis que KU est le marqueur de l'infinitif en kisongye (inf_k). Aucun verbe du kisongye (V_k) ne peut occuper la place que Y occupe. Ainsi, la relation KUKITA— 2 → X dans laquelle X serait un V_k n'est pas admise en kisongye.

Il se dégage en outre la règle d'interface kisongye-français suivante : Le verbe infinitif du français ($Vinf_f$) accepte des préfixes verbaux du kisongye mais pas les suffixes. Par exemple, pour l'énoncé (4) ci-dessus, on ne peut pas obtenir dans l'alternance de codes : « *Natravaillesha baélèves » Je travailler-CAUS. les élèves. (Je fais travailler les élèves). Il s'agit ici d'une contrainte topologique suivant laquelle Vinf_f accepte des affixes à gauche et pas à droite. Ainsi se dessine l'équivalence syntaxique suivante :

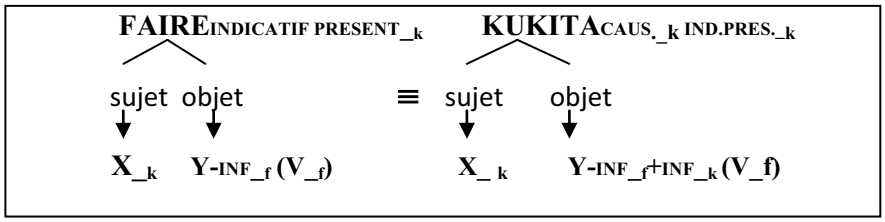


Figure 1 : SSyntS équivalente de la causation dans l'alternance de codes kisongye/français

5.2 Le verbe auxiliaire dans la structure impositive.

Le kisongye recourt au suffixe impositif « - ik -» alors que le français passe par la forme verbale pronominale :

(6) a. Meemá taáatóméká : taáatoméka : °taá – a - tom – ik – a ('se boire')
L'eau NEG il boire-IMPOS. PRES.
(L'eau n'est pas potable)

L'alternance de codes kisongye/français recourt à une construction spécifique qui n'appartient pas au lexique du kisongye bien qu'elle obéisse aux règles de formation des unités lexicales et aux règles grammaticales du cette langue. Il s'agit en fait du verbe auxiliaire « kúkitá » ('faire') qui régit un verbe pronominal du français au moyen de la translation du verbe en nom. A cet effet, le verbe infinitif occupe une position nominale :

b. Meemá taáakící kú se boire nyá
(L'eau ne faire de se boire pas)
'L'eau ne se boit pas'.

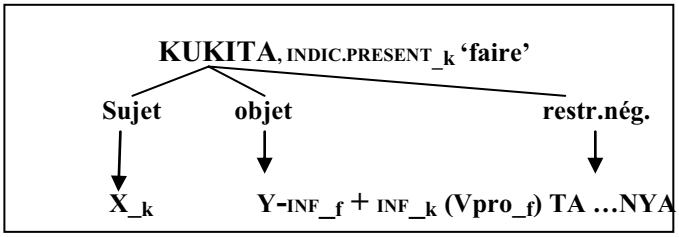


Figure 2 : SSyntS de 3b.

5.3 Le verbe auxiliaire dans une structure réciproque

L'on sait que la structure réciproque du kisongye s'exprime par le suffixe « - an -» là où le français emploie le pronom variable « se ». Il y a lieu de noter que pour les énoncés kisongye à verbes réciproques, le suffixe précité voit de plus en plus son sens 'réciproque' faiblir. D'où son renforcement au moyen du pronom réfléchi « - i -» ou « -idi - » ('se') occupant une position pré-radical :

(7) Bantu ábeesépéénáá : °a – ba – i – sep – an – a
CL2 hommes 3PL PRES.-se moquer-RECIPROQUE-PRES.
('Les hommes se moquent les uns les autres').
(8) Bantu ábasépáná : ° a – ba – sep – an – a

CL2 hommes 3PL PRES- rire-RECIPROQUE-PRES.
('Les hommes se moquent des autres').

La présence du pronom « - i - » (cf.7) apporte la nuance de réciprocité tandis que son absence implique aussi l'absence de réciprocité (cf.8). Dans l'alternance de codes kisongye/français, on assiste à l'équivalence suivante :

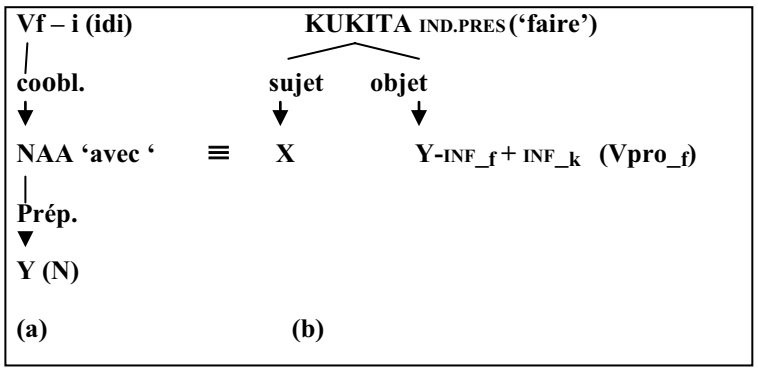


Figure 3 : SSyntS équivalentes d'une structure de réciprocité de l'alternance de codes :

La première paraphrase (cf. a) fait fonctionner le suffixe kisongye « -i -» ou « -idi -» préposé au verbe français (V_f) X. Celui-ci gouverne un nom (Y) par le biais de la préposition « NAA » ('avec'). L'on peut en déduire la règle de correspondance syntaxique de surface suivante :

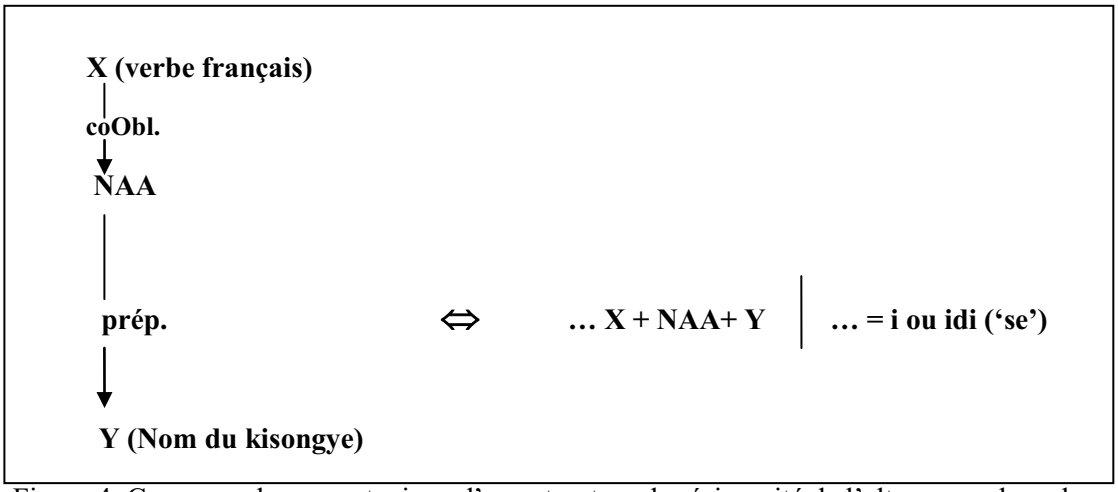


Figure 4: Correspondance syntaxique d'une structure de réciprocité de l'alternance de codes

La seconde paraphrase (cf. b) recourt au procédé déjà rencontré dans 3b s'agissant de l'impositif. Il s'agit du verbe auxiliaire « kúkitá » ('faire') gouvernant un verbe pronominal du français à l'infinitif ($Y\text{-}inf_f$) ($Vpro_f$) au moyen du translatif inf_k « KU » faisant occuper au verbe une place nominale. Les énoncés 6a et 6b ci-dessous attestent les deux structures paraphrastiques de la structure de réciprocité :

(9) a. Kasongo <u>béédivoir</u> naá professeur eetú : bá – idi - voir
(Kasongo 3SG RECIPROQUE-voir avec professeur notre) :
'Kasongo a rencontré notre professeur'
b. Kasongo <u>bákící kú se voir</u> naá professeur eetú : ba- kit – i ku se voir
(Kasongo 3SG faire-PASSE de se voir avec professeur notre) :
'Kasongo a rencontré notre professeur'.

5.4 Le verbe auxiliaire dans quelques formes verbales de l'alternance de codes.

Il est important de rappeler que le verbe auxiliaire KUKITA peut prendre, dans l'alternance de codes, toutes les flexions du kisongye s'exprimant par la suffixation, lesquelles flexions ne peuvent affecter le verbe infinitif français (V-INF_f) (cf. contrainte topologique posée dans (5.1.) ci-dessus). Présentement, il s'agit des marques du passé, de l'infinitif habituel, de l'impératif et du subjonctif. Il importe de noter également que la construction à verbe auxiliaire KUKITA est la seule qui fonctionne pour les flexions suffixées. Elle ne relève pas du kisongye et est spécifique aux V- INF_f en alternance de code. Montrons-le ci-dessous.

Le verbe auxiliaire dans la structure du passé.

Le passé s'exprime généralement par le suffixe « – ilé » du kisongye. La suffixation n'étant pas de mise dans l'alternance kisongye/français, elle est rendue de la manière suivante : si X est un V et G un grammème du kisongye qui se réalise par un suffixe, on a X_G si X est V_k et KUKITA_G si X est un Vinf._f.

Prenons à titre d'exemple, l'énoncé ci-après :

(10) Maître eetú bákícíne kuboire meemá.
(Maître notre il faire-PASSE de boire l'eau) :
'Notre maître avait bu de l'eau'.
'Mulongyeshi eetú bátóméne meemá'.

L'énoncé (7) ci-dessus se structure comme suit au niveau syntaxique de surface :

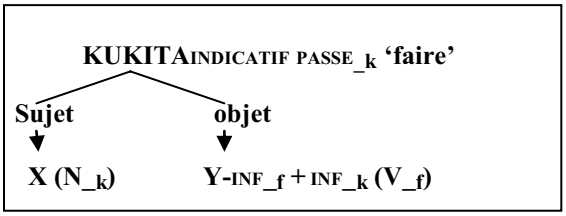


Figure 5 : SSyntS de (7).

On obtient ainsi les équivalences syntaxiques suivantes lorsqu'on passe du kisongye à l'alternance en passant par le français :

Kisongye	Français	Alternance kisongye/français
	V_f IND PASSE | aux. ↓	
V_k INDIC PASSE | objet ↓ Y (N)	V PART.PAS. | objet ↓ Y (N)	KUKITA INDICATIF PASSE_k 'faire' | objet ↓ Y-INF_f +INF_k (V-INF_f)
Bátóméne meemá	**avait bu de l'eau**	**Bákícíne kuboire meemá**

Figure 6 : SSyntS équivalentes du passé dans les trois codes

Le verbe « KUKITA » (‘faire’) gouverne l’infinitif français par le biais de la translation du verbe en nom, le translatif étant le préfixe « KU ». Cette structure est typique de l’alternance de codes.

Formes infinitives et impératives.

Nous ne parlons ici que de l’infinitif habituel qui seul admet la suffixation. Il s’obtient dans l’alternance de codes par le recours au verbe auxiliaire « KUKITA » (‘faire’) :

(11) Abikyébé kúkitánga kú travailler. : kú – kit – áng - a
(Il falloir INF.-faire habituellement INF.-de travailler) :
‘Il faut travailler habituellement’.

La SSyntS de (8) est identique à celle de la figure (5) ci-dessus, la différence étant que le verbe **auxiliaire** endosse ici les marques de l’infinitif habituel. Ce qui est dit de l’infinitif se dit aussi de l’impératif :

(12) Kítá ku dégager ndundo
(Fais de dégager la balle) :
‘Dégage la balle’.

Cet énoncé reçoit la structure syntaxique de surface suivante :

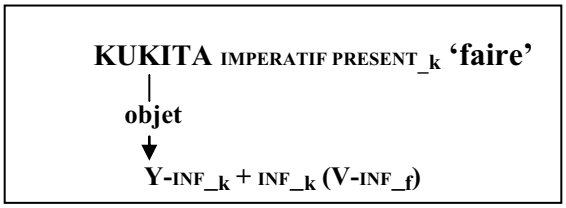


Figure 7 : SSyntS de (9).

Le subjonctif.

L’alternance de codes kisongye/français recourt au verbe auxiliaire « KUKITA » dans les mêmes conditions que celles décrites ci-dessus pour l’impératif pour exprimer le subjonctif.

Tout ce qui précède met en exergue le rôle du verbe auxiliaire « KUKITA » qui se résume dans la représentation suivante : KUKITA — objet → Y $-_{INF\_f}$ $+_{INF\_k}$ (V_f) dans laquelle KUKITA, terme gouverneur, est le verbe auxiliaire tandis que Y est un verbe français à l’infinitif.

6 Conclusion.

L’alternance de codes kisongye/français atteste le verbe auxiliaire « KUKITA » (‘faire’) ayant un rôle supplétif. Ce verbe permet de résoudre le problème de la suffixation non de mise dans l’alternance kisongye/français. En effet, lors de la synthèse des structures causatives, impositive, réciproques et lors de la synthèse des formes infinitives (infinitif habituel), de l’indicatif passé, de l’impératif et du subjonctif, le verbe auxiliaire gouverne une forme verbale française infinitive, celle-ci étant une forme translatée. La relation mettant le verbe auxiliaire comme gouverneur dans une alternance de codes est typique de celle-ci. Elle ne relève ni du français ni du kisongye.

Remerciements

Je remercie vivement Sylvain Kahane et les deux relecteurs de MTT'09 dont les remarques et suggestions ont enrichi et amélioré la qualité de ce texte.

Références.

Blanchet Philippe. 2004. « L'identification sociolinguistique des langues et des variétés linguistiques : une analyse complexe du processus de catégorisation fonctionnelle », in *MDL*, 31-36.

Gerdes Kim & Sylvain Kahane.2001. « Une grammaire TAG comme une grammaire Sens-Texte précompilée », in *TALN*, 10 -12 juin 1998, 101-111.

Gerdes Kim & Sylvain Kahane.2004. « L'amas verbal au cœur d'une modélisation topologique du français » in *K.Gerdes et C. Muller (éd.) Ordre des mots et topologie de la phrase française, Linguisticae investigationes,* n° 29, 1-14.

Guthrie M. 1970. *Comparative Bantu*, Vol.3, London, Greeg International Publisher.

Hamers J.F. & M. Blanc. 1983. *Bilingualité et bilinguisme*, Bruxelles, P. Mardaga.

Joshi A.K. 1987. « Some Problems in Processing Sentences with Intrasentential Codeswitching », in *National Language Passing ...* Cambridge University Press, 190-205.

Kahane Sylvain & Igor Mel'čuk. 2006. « Les sémantèmes de causation en français », in *La cause : approche pluridisciplinaire, Lieux*, n° 54, 489-507.

Kahane Sylvain & François Lareau. 2005. « Grammaire d'Unification Sens-Texte. Modularité et polarisation », in *TALN 2005*, 6 – 1O juin 2005.

Lüdi G. 1995. « Parler bilingue et traitement cognitif » in *Intellectica*, n° 20, 139-156.

Mel'čuk Igor. 1997. *Vers une linguistique Sens –Texte. Leçon inaugurale*, Paris, Collège de France.

Mondada L. 2007. « Code–switching comme ressource pour l'organisation de la parole en interaction », in *Journal of Language Contact, Théma 1,*168-197.

Myers-Scotton C. 1993a. *Social Motivation for Codeswitching. Evidence from Africa*, Oxford, Oxford University press.

Myers-Scotton C. 1993b. *Duelling Languages Grammatical Structures in Codeswitching*, Oxford, Clarendon Press, Oxford.

Polguère A. 1998. « La théorie Sens-Texte » in *Dialangue,* vol.8-9, 9-30.

Le temps verbal dans l'interface sémantique-syntaxe du français

François Lareau
OLST, Université de Montréal
Lattice, Université Paris Diderot (Paris 7)
francois.lareau@umontreal.ca

Abstract

We briefly introduce a new model of French verbal tenses, where two complementary inflectional categories, tense and shifting, work together to locate facts in time. The category of shifting poses a reference point in the past or in the non-past. The category of tense locates facts in relation with this reference point. The core of this article is dedicated to the modeling of tense and shifting grammemes in the semantics-syntax interface of the Meaning-Text Unification Grammar of French. We also discuss the question of grammatical idioms and collocations.

1 Introduction

Cet article aurait dû être une présentation de notre thèse de doctorat, soutenue en 2008, dans laquelle nous élaborions un nouveau modèle descriptif des temps verbaux en français. Toutefois, nous avons découvert peu après la soutenance un article de Vet (2007) qui dit essentiellement la même chose. D'un côté, la situation est frustrante, puisque nous voyons la paternité de ce modèle nous échapper de peu. Mais voyons le bon côté des choses : l'article que nous voulions vous présenter ici est déjà écrit, ce qui nous épargne beaucoup de travail et nous permettra de nous concentrer plutôt sur la formulation de ce modèle dans le cadre de la Grammaire d'Unification Sens-Texte (GUST).

Nous allons d'abord présenter brièvement à la section 2 l'hypothèse de Vet (2007) et de Lareau (2008), avant de montrer comment elle se modélise en GUST à la section 3.

2 Une nouvelle approche reichenbachienne

Partant de prémisses différentes, Vet (2007) et Lareau (2008) arrivent indépendamment aux mêmes conclusions quant à l'organisation du système temporel verbal du français.

2.1 L'inadéquation du modèle de Reichenbach

Les travaux de Vet se situent dans le cadre du discours et de la sémantique formelle. Dans son article (2007), il part du constat que le modèle de Reichenbach (1947), développé d'abord pour l'anglais mais prétendant à l'universalité, n'est pas adéquat pour la description du français. La thèse principale de Reichenbach était que les temps verbaux n'expriment pas forcément des relations temporelles entre le fait dénoté par le verbe et le moment d'énonciation, mais que peut intervenir un troisième moment. Les faits, plutôt que d'être directement situés par rapport au moment d'énonciation, le sont par rapport à ce point de référence, qui lui-même l'est par rapport au moment d'énonciation. En représentant le moment où se déroule le fait par E, le moment d'énonciation par S et le point de référence par R, Reichenbach propose le modèle suivant[1] :

[1] La relation entre E et S n'est pas pertinente dans ce modèle puisque E n'est jamais directement situé par rapport à S.

TAB. 1 – Le modèle de Reichenbach

	Passé ($R < S$)	**Présent** ($R = S$)	**Futur** ($R > S$)
Antérieur ($E < R$)	*He had eaten*	*He has eaten*	*He will have eaten*
Simple ($E = R$)	*He ate*	*He eats*	*He will eat*
Postérieur ($E > R$)	*He would eat*	*He will eat*	—

Or, ce modèle ne suffit pas pour la description du français, notamment, comme le note Vet (2007), parce que les temps du passé dans cette langue sont plus nombreux que ceux du futur, une réalité que ne permet pas de saisir le modèle parfaitement symétrique de Reichenbach. Qui plus est, le « futur postérieur » prévu par ce modèle ne serait pas attesté dans les langues naturelles, selon Vet.

2.2 Une méthodologie basée sur la lexicologie explicative et combinatoire

Dans notre thèse (Lareau, 2008), nous partions plutôt du constat que les modèles existants ne décrivent pas toujours les signes dans leur ensemble, mais ne s'intéressent souvent qu'à une seule de leurs composantes (sens, forme ou combinatoire). En nous inspirant de la lexicologie explicative et combinatoire (Mel'čuk et al., 1995), nous y proposions une méthodologie pour la description des grammèmes que nous résumons ici.

Nous concevons le grammème non pas comme un signe, mais comme une entité du niveau d'abstraction du vocable dont il faut distinguer les différentes acceptions, que nous appelons « grammies »[2]. Parmi les grammies associées à un grammème, une est considérée comme la grammie de base. On peut l'identifier grâce à un ensemble de critères basés sur le sens et la combinatoire des signes en jeu (Lareau, 2008, pp. 53–60). C'est la grammie de base des grammèmes qui guide la construction du modèle linguistique. C'est en effet en fonction du sens et de la combinatoire de celle-ci que les grammèmes sont regroupés en catégories flexionnelles. Les grammies de base des grammèmes d'une même catégorie flexionnelle doivent être mutuellement exclusives et avoir une combinatoire similaire, en plus de présenter une parenté sémantique évidente (Lareau, 2008, pp. 62–67). Il faut toutefois considérer que les grammèmes peuvent se phraséologiser (Beck, 2007), ce qui peut brouiller les pistes ; nous y reviendrons à la section 3.2.

Quoi qu'il en soit, malgré que nos prémisses, notre méthodologie et notre cadre théorique diffèrent de ceux de Vet (2007), nous arrivons aux mêmes conclusions, que nous résumons ci-dessous.

2.3 La polysémie du passé composé

Il est généralement admis, au moins depuis Benveniste (1959), que le passé composé peut exprimer soit l'antériorité (il commute alors aisément avec le passé simple et est compatible avec un complément « *en* + durée »), soit l'aspect accompli (auquel cas la substitution n'est pas possible, et on peut utiliser un complément « *depuis* + durée ») :

(1) a. J'ai terminé le tableau en deux heures.

b. Je terminai le tableau en deux heures.

(2) a. J'ai terminé le tableau depuis deux heures.

b. * Je terminai le tableau depuis deux heures.

Il y a donc clairement deux acceptions au passé composé. On les retrouve d'ailleurs toutes les deux combinées dans les formes surcomposées, où l'état dénoté par l'accompli est situé dans le passé (Vet, 2007;

[2]Le terme vient de Kahane (2002), qui observait que les grammies sont des signes profonds au même titre que les lexies. En ce sens, le signifiant d'une grammie est un grammème, de la même façon que le signifiant d'une lexie est un lexème. L'article de Kahane (2009) dans le présent volume revient sur ce thème. Précisons que les grammies, comme les lexies, sont des patrons ou regroupements de signes. Cependant, dans cet article, nous les traiterons comme des signes afin d'alléger le texte.

Lareau, 2008). Nous appellerons la lexie qui exprime l'antériorité AVOIR$_{\text{ant}}$, et celle qui exprime l'accompli, AVOIR$_{\text{acc}}$. Dans ce qui suit, nous allons laisser de côté cette dernière puisqu'elle ne fait pas partie des marqueurs de temps, mais plutôt de ce que nous appelons les marqueurs de phase, au même titre que ALLER (Vet, 2007; Lareau, 2008).

2.4 Un système temporel à deux catégories flexionnelles complémentaires

Les formes verbales à valeur temporelle du français peuvent être divisées en deux classes : celles qui situent les faits directement par rapport au moment d'énonciation (le présent, le passé composé d'antériorité, le passé simple et le futur simple) et celles qui situent les faits par rapport à un repère temporel qui se trouve avant le moment d'énonciation (l'imparfait, le plus-que-parfait d'antériorité et le conditionnel à valeur temporelle)[3]. Le sens de base de ces formes peut être représenté en termes de deux séries de trois sens : 'simultané' $\sim$ 'antérieur' $\sim$ 'postérieur' d'une part et 'par rapport à un repère passé $\sim$ non-passé' d'autre part, ce que nous représentons sous forme de tableau ci-dessous[4] :

TAB. 2 – Un découpage sémantique des temps verbaux du français

	$T_1 \geqslant T_0$	$T_1 < T_0$
$T < T_1$	*mangea / a mangé*	*avait mangé*
$T \approx T_1$	*mange*	*mangeait*
$T > T_1$	*mangera*	*mangerait*

Or, ce découpage sémantique va très clairement de paire avec un découpage formel (sauf pour le passé simple, ce qui n'est pas gênant étant donné son usage restreint au registre littéraire). Nous avons donc affaire à cinq grammies, qui sont les acceptions de base d'autant de grammèmes que nous regroupons en deux catégories flexionnelles. D'abord, la catégorie de décalage, dont les grammèmes, dans leur acception de base, situent un point de repère par rapport au moment d'énonciation :

- **Non-décalé** signifie, dans son acception de base, que le point de repère est soit le moment d'énonciation, soit un moment dans le futur[5]. Il s'exprime par un suffixe zéro.
- **Décalé**, dans son acception de base, indique que le repère temporel est dans le passé. Il s'exprime par le suffixe –AI– (dont les allomorphes sont *–ai–* [*faisait*] et *–i–* [*faisions*]).

Ensuite, la catégorie flexionnelle de temps, qui contient trois grammèmes qui servent, dans leur acception de base, à situer les faits par rapport au point de repère :

- Le grammème **antérieur**[6], dont la grammie de base signifie 'X a lieu avant Y'. Il s'exprime généralement de façon analytique par la construction « *avoir*$_{\text{ant}}$ + V–*é* ». Dans le cas du passé simple, le grammème **antérieur** n'a pas de signifiant propre et se trouve exprimé dans le même morphe que la personne et le nombre.
- Le grammème **simultané**, dont la grammie de base signifie 'X a lieu en même temps que Y'. Il s'exprime par un signe zéro, à savoir l'absence d'auxiliaire et de suffixe de temps.
- Le grammème **postérieur**, dont la grammie de base signifie 'X a lieu après Y'. Il s'exprime de façon synthétique par le morphème –R– (le suffixe du futur simple).

[3] Cette dichotomie entre temps absolus et temps relatifs en français est connue au moins depuis le XVIII$^{\text{e}}$ siècle. Voir notamment Weinrich (1973) et Comrie (1985).

[4] T représente le moment où le fait dénoté par le verbe se produit, T_0, le moment d'énonciation et T_1, le repère par rapport auquel est situé T. Ils correspondent respectivement aux R, E_0 et P de Vet (2007). Nos T et T_0 correspondent, respectivement, aux E et S de Reichenbach. Touteois, notre T_1 diffère de son R, qui est problématique (Molendijk, 1990).

[5] Le français ne permet pas de situer explicitement un fait par rapport à un repère futur. Voir Lareau (2008, pp. 208–210).

[6] Nous voudrions appeler les grammèmes de temps « passé », « présent » et « futur », puisqu'ils correspondent à peu près à ceux qu'on trouve dans beaucoup de langues et que cette terminologie est courante. Cependant, puisque nous utilisons les termes traditionnels pour nommer les formes verbales (« présent », « imparfait », etc.), le risque de confusion serait trop grand.

Les catégories de temps et de décalage sont toutes deux obligatoires à l'indicatif. Il n'y a donc aucune forme verbale qui exprime, par exemple, uniquement le grammème **antérieur** ou uniquement le grammème **décalé**, ce qui diffère des modèles comme celui de Martinet (1979), où l'imparfait n'exprime que le « monème » **décalé**. Dans notre modèle, l'imparfait n'est pas indécomposable, mais est formé de la combinaison de deux grammèmes : **simultané** et **décalé**. Ainsi, nous avons pour le français un système temporel à deux dimensions qui diffère aussi des analyses du type « temps absolus *vs* temps relatifs » ou d'autres systèmes à paradigmes concurrents, comme ceux de Imbs (1960) et de Weinrich (1973), ou l'ancien modèle de Vet (1980), en ce que les temps de l'indicatif expriment toujours deux grammèmes. Vet (2007) ne formule pas son nouveau modèle explicitement en termes de grammèmes comme nous le faisons, mais son analyse des faits est la même.

TAB. 3 – Un système temporel à deux catégories flexionnelles complémentaires

		Décalage	
		Non-décalé	**Décalé**
Temps	**Antérieur**	*mangea* / *a mangé*	*avait mangé*
	Simultané	*mange*	*mangeait*
	Postérieur	*mangera*	*mangerait*

Ce modèle rend compte de façon élégante notamment de la polysémie parallèle de certaines paires de formes comme le présent et l'imparfait (de validité permanente, d'habitude, de passé récent, de futur inéluctable, historiques ou de condition) ou encore le futur simple et le conditionnel (historique ou d'atténuation). Considérons le cas du présent et de l'imparfait, qui présentent d'importantes similarités de sens. Labeau (2002) a recensé les valeurs de l'imparfait dans cinq grammaires récentes. Nous avons fait la même chose avec le présent dans le *Bescherelle* (1998, § 143), *Le bon usage* (1993, § 850) et la grammaire de Riegel, Pellat & Rioul (1994, pp. 299–301). Le Tableau 4 ci-dessous résume les principales acceptions de ces deux formes.

TAB. 4 – Les principaux sens du présent et de l'imparfait dans quelques grammaires de référence

Acceptions	**Présent**	**Imparfait**
Actualité	*Il pleut.*	*Il pleuvait.*
Validité permanente	*Le soleil se couche à l'Ouest.*	*Il savait que le soleil se couchait à l'Ouest.*
Habitude	*Il mange quatre fois par jour.*	*Il mangeait quatre fois par jour à l'époque.*
Passé récent	*Je rentre tout juste de Berlin*	*Je rentrais tout juste de Berlin*
Futur inéluctable	*Nous partons à 5h demain.*	*Nous partions à 5h le lendemain.*
Historique	*Le 2 octobre 1535, Cartier arrive à Hochelaga.*	*Le 2 octobre 1535, Cartier arrivait à Hochelaga.*
Conditionnel	*S'il pleut, on ira au cinéma.*	*Elle a dit que s'il pleuvait, on irait au cinéma.*
Injonctif	*On se calme !*	—
Atténuation	—	*Je voulais vous demander quelque chose.*
Irréel	—	*Si j'étais riche, je partirais en voyage autour du monde.*

Le parallèle est frappant : presque toutes les valeurs du présent trouvent leur écho à l'imparfait, et vice-versa. Cela s'explique par le fait que le présent et l'imparfait expriment tous les deux le même grammème **simultané**. Les sens parallèles de ces deux formes correspondent en fait aux différentes acceptions de ce

grammème. La différence d'ancrage temporel qu'on observe entre la série du présent et celle de l'imparfait est due au fait que le présent exprime le grammème **non-décalé**, alors que l'imparfait exprime **décalé**.

Ce modèle explique également de façon toute naturelle le phénomène de la concordance des temps. Par exemple, considérons les phrases suivantes :

(3) a. Il pense qu'elle viendra.
    b. Il pensait qu'elle viendrait.
    c. Il pensera qu'elle viendra.

(4) a. Il dit qu'elle est laide.
    b. Il a dit qu'elle était laide.
    c. Il dira qu'elle est laide.

Le verbe subordonné situe un fait par rapport au moment où a lieu le fait dénoté par le verbe principal. Il porte alors le grammème de temps qui exprime la relation en question (**postérieur** pour les trois premières phrases, **simultané** pour les trois autres). Pour ce qui est du grammème de décalage, son choix dépend du moment où se déroule le fait dénoté par le verbe principal. Si ce dernier est dans le passé, comme en (3b) et (4b), alors le verbe subordonné portera le grammème **décalé** puisqu'il est situé par rapport à un repère qui est dans le passé ; autrement, il prendra le grammème **non-décalé**.

Il faut noter que quand le verbe de la subordonnée dénote un état qui transcende les époques (comme c'est le cas en (4a–c) ci-dessus), alors le locuteur a le choix de situer cet état par rapport au référent du verbe principal ou par rapport au moment d'énonciation. Dans le cas où le verbe principal est au présent ou au futur, cela ne fait aucune différence visible, puisque c'est le même grammème **non-décalé** qui doit être utilisé, peu importe si le point de référence est actuel ou ultérieur. Par contre, dans le cas d'un verbe principal au passé, on peut observer le phénomène :

(5) a. Il a dit qu'elle était laide.
    b. Il a dit qu'elle est laide.

Il s'agit d'un choix du locuteur, selon qu'il souhaite prendre un point de vue décalé ou non, en fonction de ses buts communicatifs.

En bref, le système proposé ici nous semble adéquat et économique. Le fait que Vet (2007) et nous-même soyons arrivés aux mêmes conclusions quant à l'organisation du système flexionnel verbal du français nous conforte dans notre hypothèse. Nous allons maintenant voir comment cela se modélise en GUST.

3 Les grammèmes de temps et de décalage dans l'interface sémantique-syntaxe

Nous n'avons pas l'espace nécessaire ici pour présenter les fondements de GUST, aussi renvoyons-nous le lecteur aux travaux de Kahane (2002; 2004), Kahane & Lareau (2005a; 2005b) et Lareau (2008). Les représentations en GUST ont l'avantage d'être intuitives et faciles à décoder, donc le lecteur non averti devrait pouvoir suivre notre propos. Il faut savoir que les structures se combinent par unification et que tous les objets sont polarisés. La polarité blanche représente un manque à combler (elle correspond à peu près au « contexte » des règles de la TST classique). La polarité noire représente un objet saturé. L'unification de deux objets blancs donne un objet blanc ; un objet blanc et un noir donnent un objet noir ; deux objets noirs ne peuvent pas s'unifier. Les structures doivent s'unifier jusqu'à ce que tous les objets soient saturés, dans lequel cas on a une structure bien formée.

Nous ne nous intéresserons dans le cadre de cet article qu'à l'interface sémantique-syntaxe. Pour une discussion des règles des grammaires de bonne formation sémantique et syntaxique pertinentes pour la flexion verbale en français, ainsi que l'articulation de ces modules, voir Lareau (2008).

3.1 L'acception de base des grammèmes de temps et de décalage

Il n'y a que trois sémantèmes temporels en jeu dans la flexion verbale : 'simultané', 'antérieur' et 'postérieur'. Ils correspondent aux trois grammies de base de la catégorie flexionnelle de temps. Selon qu'ils situent les faits par rapport à un repère passé ou non, ils déclenchent aussi l'utilisation des grammèmes **décalé** ou **non-décalé** de la catégorie flexionnelle de décalage.

Dans leur acception de base, les grammèmes **simultané** et **postérieur** expriment tout simplement les prédicats 'simultané' et 'postérieur', qui situent un fait par rapport à un autre. Cela se modélise par les règles de la figure 1 ci-dessous. Puisque le point de repère peut être soit le moment d'énonciation, soit un autre fait, le nœud qui le représente ne porte pas d'étiquette dans ces règles. Il pourra alors s'unifier avec n'importe quel nœud sémantique.

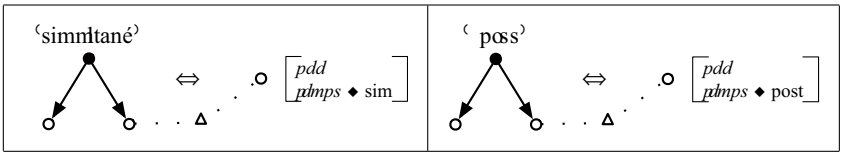


FIG. 1 – L'expression des sémantèmes 'simultané' et 'postérieur'

Le sémantème 'antérieur' peut s'exprimer en français soit par $\text{AVOIR}_{\text{ant}}$ (ou ÊTRE_{ant}), soit par le passé simple. Il y a donc trois règles concurrentes qui mettent en correspondance le même prédicat avec chacun de ces marqueurs. Les deux premières règles de la figure 2 ci-dessous introduisent un auxiliaire (selon le syntactique de l'auxilié) et lui imposent le grammème de temps **simultané**. Ils imposent aussi la finitude **participe-é**[7] au verbe qui en dépend. Ces grammèmes font partie du signifiant de la grammie ANTÉRIEUR_1. Les auxiliaires de temps forcent la montée du sujet, ce qui est représenté par une quasi-dépendance (en pointillés)[8]. La troisième règle, quant à elle, introduit simplement le grammème **antérieur**, qui pourra être mis en correspondance avec le suffixe du passé simple par la grammaire d'interface syntaxe-morphotopologie (mais seulement si l'actant de 'avant' est 'maintenant', et uniquement dans un registre littéraire).

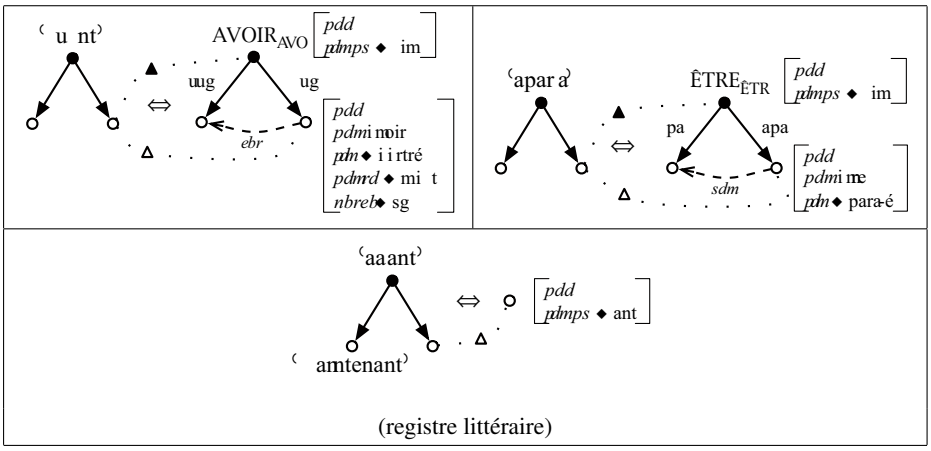


FIG. 2 – L'expression du sémantème 'antérieur'

Dans leur acception de base, les grammèmes de décalage n'expriment pas des sémantèmes en soi, mais plutôt des configurations de sémantèmes. Les règles qui les introduisent ne saturent donc aucun objet du ni-

[7] En gros, il s'agit du participe passé. Voir Lareau (2008).

[8] Voir Kahane (2002, p. 37) ou Lareau (2008, pp. 324–326)

veau sémantique[9]. Elles construisent néanmoins un objet au niveau syntaxique (le grammème de décalage). La grammie de base du grammème **non-décalé** indique que le repère temporel est le moment d'énonciation ou un autre moment dans le futur. Si, au contraire, un fait est situé temporellement par rapport à un autre fait qui lui-même se trouve avant le moment d'énonciation, alors le verbe qui exprime ce premier fait porte le grammème **décalé**. C'est ce que modélisent les règles de la figure 3 ci-dessous.

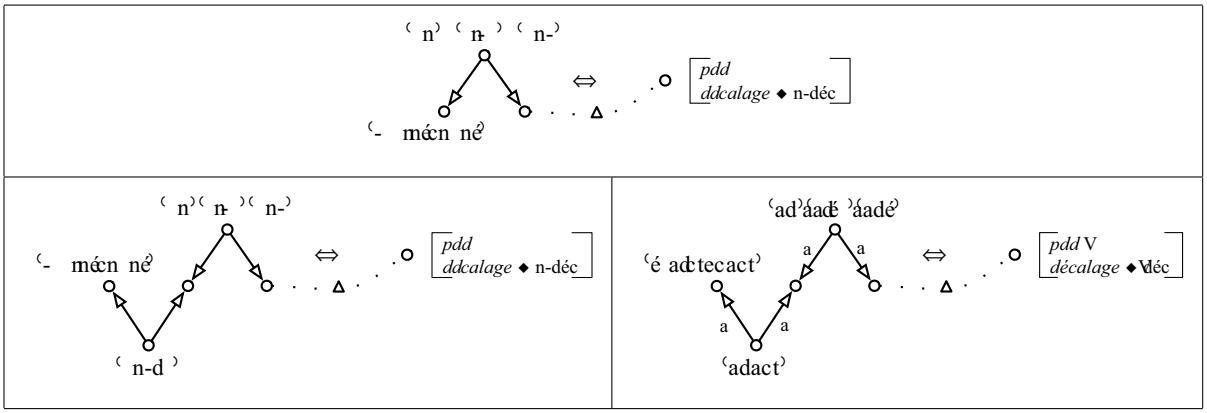


FIG. 3 – L'acception de base des grammèmes de décalage

3.2 Le traitement des locutions et collocations grammémiques

Les temps de l'indicatif en français n'expriment pas toujours des sens purement temporels. Nous allons considérer ici le cas du futur de supposition, du conditionnel de réserve et du futur antérieur rétrospectif.

Le futur simple peut signifier une supposition de la part du locuteur [*Ils tardent. Ils seront sans doute perdus*]. Ce sens est exprimé par la combinaison **postérieur** $\oplus$ **non-décalé** ; il s'agit d'une locution grammémique[10]. Elle se traite de la même façon que le conditionnel de réserve [*Le suspect aurait fait feu en direction des policiers*], une autre locution grammémique, dont le signifiant est **postérieur** $\oplus$ **décalé**. Il s'agit de locutions puisque leur sens est exprimé au niveau syntaxique par la combinaison de deux grammèmes qui, individuellement, n'expriment ni ce sens, ni une de ses composantes. Ces deux locutions se modélisent par les règles de la figure 4 ci-dessous.

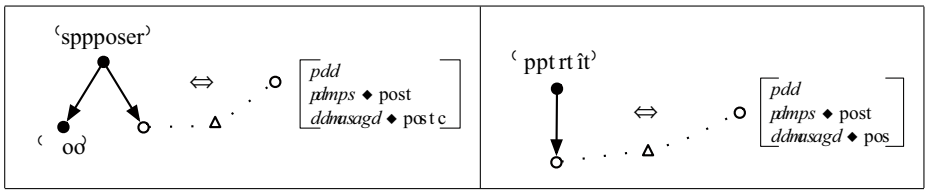


FIG. 4 – Deux locutions grammémiques

Ces locutions grammémiques sont l'équivalent, dans le domaine grammatical, des locutions lexicales comme *prendre le taureau par les cornes* ou *faire ses choux gras* : à un seul nœud sémantique correspondent plusieurs objets au niveau syntaxique. Ce qui distingue les locutions grammémiques des locutions lexicales, c'est uniquement la nature des signifiants.

[9] Un cas de figure qui n'était pas prévu par Kahane & Lareau (2005a; 2005b).

[10] Beck (2007) suggérait le terme de « phrasème morphologique ». Cependant, puisque nous ne traitons pas le niveau morphologique, nous préférons parler de locutions grammémiques.

On sait que, dans le domaine lexical, il existe d'autres types d'expressions phraséologisées, notamment les collocations (par exemple, *froid de canard* ou *remporter la victoire*). Or, nous allons voir que les collocations existent également dans le domaine grammatical. Considérons la phrase suivante :

(6) L'hiver 2008–2009 aura été particulièrement froid.

Dans cette phrase, le sens de « *X aura* V–*é* » est plus ou moins 'en rétrospective, *X* a V–é'. Il contient donc le sens de AVOIR$_{acc}$ (le marqueur de l'accompli), mais pas celui du grammème **postérieur**, puisque la phase dénotée par l'accompli n'est aucunement située dans le futur par rapport à quelque point de référence que ce soit. Il s'agit d'un cas intéressant de phraséologie, qu'on pourrait appeler « collocation grammémique ». Cette notion, à notre connaissance, n'a jamais été discutée dans la TST (ni dans d'autres cadres théoriques, pour autant que nous sachions). Nous allons donc commencer par montrer comment se formalisent les collocations lexicales ; nous pourrons ensuite mieux apprécier les parallèles avec cette collocation grammémique.

Dans le cadre de la TST, les collocations sont décrites dans le dictionnaire par des fonctions lexicales (Mel'čuk et al., 1995). Or, en GUST, il n'y a pas de dictionnaire : tous les signes sont décrits par des règles. Kahane & Polguère (2001) ont démontré que les fonctions lexicales peuvent être définies formellement par des patrons de correspondance entre des fragments de structures sémantiques et syntaxiques. Il est donc possible de les représenter par des règles de correspondance. Prenons par exemple la collocation *peur bleue*. Son sens est 'peur intense'. Dans cette expression, le lexème PEUR exprime son sens habituel ; il s'agit de la base de la collocation. Par contre, BLEU ne signifie 'intense' que dans le contexte de PEUR (et ce sens s'exprime de préférence par ce lexème dans ce contexte). Le lexème BLEU n'a donc pas été choisi tout à fait librement. C'est le collocatif de la collocation. La figure 5 ci-dessous montre, à gauche, la règle qui formalise cette lexicalisation contrainte. Il faut la comparer avec la règle de droite, qui modélise la lexicalisation, moins contrainte, de 'intense' par l'adjectif INTENSE. Cette dernière ne tient pas compte du contexte lexical, alors que la règle de gauche, qui décrit la collocation *peur bleue*, indique que 'intense' ne peut se lexicaliser par BLEU que si cet adjectif modifie PEUR[11].

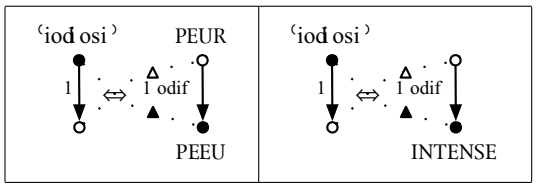


FIG. 5 – Collocation *vs* lexicalisation libre dans le domaine lexical

Le cas du futur antérieur rétrospectif qu'on trouve en (6) est très similaire. Son sens inclut celui de AVOIR$_{acc}$: 'l'année aura$_{acc}$ été froide' = 'en rétrospective, l'année a$_{acc}$ été froide'. La phase accomplie est la base de la collocation. Le sens 'en rétrospective' ne s'exprime habituellement pas par la combinaison des grammèmes **postérieur** et **non-décalé** comme en (6). Ce n'est que dans le contexte de l'accompli que cette combinaison de grammèmes exprime ce sens. Ils forment ensemble le collocatif. Formellement, ce signe se décrit par la règle à la figure 6.

Contrairement à la règle qui décrit la collocation *peur bleue* à la figure 5, où le sémantème correspondant à la base n'était pas pertinent, ici c'est le sémantème qui importe. En effet, le sens de l'accompli (que nous avons représenté ici par 'avoir$_{acc}$') peut s'exprimer soit par l'auxiliaire AVOIR$_{acc}$, soit par ÊTRE$_{acc}$, ce qui n'a pas d'incidence sur la collocation :

(7) a. L'année dernière aura été décevante.

[11] Le sémantème qu'exprime PEUR n'est pas pertinent dans cette règle. Ce qui compte, c'est le lexème, c'est pourquoi le sémantème n'est pas étiqueté. Rappelons que seuls les objets polarisés en noir sont « construits » par la règle.

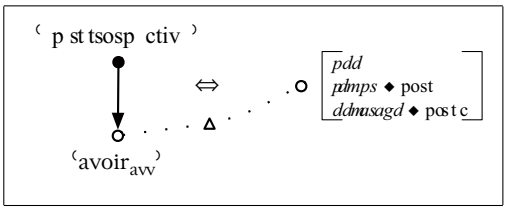


FIG. 6 – Le futur antérieur rétrospectif : une collocation grammémique

b. L'année dernière se sera terminée comme elle avait commencé.

Le fait que la collocation grammémique de la figure 6 soit contrôlée par un sémantème, et non un élément de la structure syntaxique comme à la figure 5, n'est pas gênant. On retrouve ce phénomène dans le domaine lexical également. Milićević (1997) montre en effet que les étiquettes sémantiques peuvent contrôler des collocations. Elle donne notamment l'exemple $Func_0$('événement') = *survenir*, *avoir lieu*, *se produire*. Toutes les lexies qui dénotent des événements héritent de cette cooccurrence lexicale, par exemple : *L'explosion s'est produite <a eu lieu, est survenue> à 3 h*.

4 Conclusion

Dans notre thèse de doctorat, nous somme arrivé, de façon indépendante, à un modèle des temps de l'indicatif en français pratiquement identique à celui de Vet (2007). Nous avons présenté ici ce modèle, où les faits sont situés temporellement par le truchement de deux catégories flexionnelles complémentaires. La première catégorie, celle de décalage, indique si le point de repère est décalé dans le passé ou non. La seconde, celle de temps à proprement parlé, situe le fait dénoté par le verbe par rapport au point de référence en question. Aucune forme verbale ne porte seulement qu'un grammème de temps ou de décalage ; ces deux catégories flexionnelles fonctionnent toujours ensemble.

Nous avons brièvement exposé la méthodologie qui nous a permis d'arriver à ces conclusions. Elle repose sur une conception du grammème comme une entité du niveau d'abstraction du vocable. Un même grammème peut avoir plusieurs acceptions, des grammies, dont une est l'acception de base. C'est la grammie de base des grammèmes qui guide la description du système flexionnel.

Nous avons vu que la description formelle des grammèmes temporels du français dans l'interface sémantique-syntaxe de GUST était simple et intuitive, ce qui démontre l'utilité de ce formalisme. Nous avons remarqué que les grammèmes de décalage avaient la particularité de ne saturer aucun objet au niveau sémantique, ce qui était inattendu.

Nous avons également montré comment se traitent les phrasèmes grammémiques en GUST. Nous avons vu d'abord que les locutions grammémiques que sont le conditionnel de réserve et le futur de supposition se décrivent de la même façon que les locutions lexicales, si ce n'est de la nature des signifiants qui diffère. Nous avons également démontré l'existence de collocations grammémiques en donnant l'exemple du futur antérieur rétrospectif, et nous avons vu que ces collocations se décrivent de façon similaire aux collocations lexicales.

Lors d'une expérience dans le cadre du projet *Marquis*[12], un générateur de textes multilingue, nous avons réutilisé notre modèle pour le catalan, l'espagnol et le portugais. Nous avons constaté que ces langues montraient la même organisation à ce niveau. Nous avons toutefois noté que la sélection du point de repère ne se faisait pas forcément de manière identique dans toutes ces langues. Par exemple, l'imparfait dans *Ce matin, la concentration d'ozone <u>était</u> de 25 µg/m³* se traduit par le passé *va ser* en catalan, et non par l'imparfait **era*. Ce fait suggère que la saillance communicative des circonstanciels n'est pas la même d'une langue à l'autre, et que cela influence le choix du point de repère temporel. Il semble donc que le squelette de notre

[12] Voir Wanner & Lareau (2009).

modèle ait une utilité pour la description d'autres langues, mais que de d'autres phénomènes linguistiques soient à l'œuvre, ce qui appelle de nouvelles recherches.

Remerciements

Nous remercions les deux lecteurs anonymes pour leurs commentaires, qui nous ont permis d'améliorer ce texte.

References

Beck, David. 2007. Morphological phrasemes in totonacan inflection. In *Proceedings of MTT 2007*, Klagenfurt.

Benveniste, Émile. 1959. Les relations de temps dans le verbe français. *Bulletin de la Société de linguistique de Paris*, 54(1) :69–82.

Bescherelle. 1998. *Bescherelle. L'art de conjuguer : dictionnaire de 12 000 verbes*. Hurtubise HMH, Montréal.

Comrie, Bernard. 1985. *Tense*. Cambridge University Press, Cambridge.

Grevisse, Maurice & André Goosse. 1993. *Le bon usage : grammaire française*. Duculot, Paris, 13e edition.

Imbs, Paul. 1960. *L'emploi des temps verbaux en français moderne*. Klincksieck, Paris.

Kahane, Sylvain & François Lareau. 2005a. Grammaire d'unification sens-texte : modularité et polarisation. In *Actes de TALN 2005*, pages 23–32, Dourdan.

Kahane, Sylvain & François Lareau. 2005b. Meaning-Text unification grammar : modularity and polarization. In *Proceedings of MTT 2005*, pages 163–173, Moscou.

Kahane, Sylvain & Alain Polguère. 2001. Formal foundation of lexical functions. In *Proceedings of ACL 2001*, Toulouse.

Kahane, Sylvain. 2002. *Grammaire d'Unification Sens-Texte : Vers un modèle mathématique articulé de la langue*. Document de synthèse pour l'habilitation à diriger les recherches, Université Paris 7.

Kahane, Sylvain. 2004. Grammaires d'unification polarisée. In *Actes de TALN 2004*, Fès.

Kahane, Sylvain. 2009. Defining the deep syntactic structure : How the significant units combine. In *Proceedings of MTT 2009*, Montréal.

Labeau, Emmanuelle. 2002. L'unité de l'imparfait : vues théoriques et perspectives pour les apprenants du français, langue étrangère. *Travaux de linguistique*, 45 :157–184.

Lareau, François. 2008. *Vers une grammaire d'unification Sens-Texte du français : le temps verbal dans l'interface sémantique-syntaxe*. Thèse de doctorat, Université de Montréal / Université Paris 7.

Martinet, André. 1979. *Grammaire fonctionnelle du français*. CREDIF, Paris.

Mel'čuk, Igor Aleksandrovič, André Clas, & Alain Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*. Universités francophones. Duculot, Louvain-la-Neuve.

Milićević, Jasmina. 1997. Étiquettes sémantiques dans un dictionnaire formalisé du type dictionnaire explicatif et combinatoire. Master's thesis, Université de Montréal.

Molendijk, Arie. 1990. *Le passé simple et l'imparfait : une approche reichenbachienne*. Rodopi, Amsterdam / Atlanta.

Reichenbach, Hans. 1947. *Elements of symbolic logic*. Macmillan, New York.

Riegel, Martin, Jean-Christophe Pellat, & René Rioul. 1994. *Grammaire méthodique du français*. Presses universitaires de France, Paris.

Vet, Co. 1980. *Temps, aspect et adverbes de temps en français contemporain. Essai de sémantique formelle*. Droz, Genève.

Vet, Co. 2007. The descriptive inadequacy of reichenbach's tense system : a new proposal. In de Saussure, Louis, Jacques Moeschler, & Genoveva Puskas, editors, *Tense, Mood and Aspect : Theoretical and Descriptive Issues*, pages 7–26. Rodopi, Amsterdam / New York.

Wanner, Leo & François Lareau. 2009. Applying the Meaning-Text theory model to text synthesis with low- and middle-density languages in mind. In Nirenburg, Sergei, editor, *Language Engineering for Lesser-Studied Languages*, volume 21 of *NATO Science for Peace and Security Series - D : Information and Communication Security*. IOS Press, Amsterdam.

Weinrich, Hans. 1973. *Le temps*. Le Seuil, Paris. Traduction de Tempus.

Semantic Resources for Textual Content Compression

Nina N. Leontyeva
Research Computing Center
Moscow State University
Leninskije Gory, Moscow, 119899
Russian Federation
leont-nn@yandex.ru

Abstract

We discuss an information-linguistic model designed to extract the most essential parts of the content from any text, being at the same time a model of "soft" understanding of texts. The linguistic aspects of the model and the properties of the semantic dictionary ensuring the content analysis with compression of the text structure are considered. Our dictionary resources serve an instrument for building a multilevel textual semantic structure. The primary semantic representation of a whole text may be used as the starting point for generating complex linguistic units ('Situation' and 'Textual Fact') as candidates for Events.

1 Multiple Dimensions of Textual Structure

The shortest representation of a text within an Information System is a "Search Pattern" of this text in the form of a simple list of key-words and/or terms. Any short exposition of the initial text – an abstract, a summary, an annotation, a free rendering, a paraphrase, etc., be it human or computer-aided product, – gives some dimensions to textual representation. The well-known aspect of text understanding in Information Extraction (IE) systems is "partial understanding," with different degrees of compression of the initial text content. IE-systems construct different special frames: of persons (including VIPs), organizations, geographical names, parametric or referential information, etc. (see, e.g., Grishman, 1999 and materials of MUC, 1-7). IE-systems that create various TASK- or USER-oriented frame structures often deal with very simple linguistic data. (Mani, 2001) regards those systems as different means of text compression: "condensation of document information content for the benefit of the reader and task." These and similar sets of specific databases (DBs) form several "content dimensions" of a text.

On the other hand, we can obtain the most detailed representation of a text as a result of linguistic analysis. First of all, theoretical linguistics does not allow for any loss of information expressed by the text; the same is true for Semantic structures or representations (SemS, SemR), being elaborated in Meaning-Text Theory (MTT). Systems of NLP-analysis based on MTT try to retain even the smallest portion of meaning expressed in a sentence, which is very important, especially for Machine Translation (e.g., the ETAP system). The whole text semantic structure appears in ETAP as a sequence of sentence SemRs. It is a special kind of understanding – the purely linguistic one. Including a set of linguistic structures in the set of whole text understanding (WTU) structures we emphasize the principal role of linguistics in modeling intellectual systems that must rely on full linguistic analysis.

In an inverse task of Text Generation (TG), the source structure may be a table, a schema, a picture or a specific database (see McKeown, 1988 and many others issues). If some of them may serve as a basis for producing a natural text (as in TG-systems by Kittredge and Iordanskaja), then the source structures may be regarded as some compressed content of a future text, the real "textual" representations, thus enlarging the set of compressed products.

Any set of the SemRs mentioned above and of similar ones would present a kind of multidimensional semantic (or simply, **multisemantic**) structure modeling different modes of human text understanding.

The final multisemantic structure of a given text or of a corpus would serve first of all the basis for differentiated search processes. The information retrieved may be combined into some compressed semantic representation of a corpus (in our model it would be a BTF, or Base of Textual Facts, see below). This compressed product may be expanded in its turn into natural text: summary, etc.

"Text understanding" denotes a sequence of operations that extract information from any given text with the completeness degree desired by the user. It is a central point of the proposed **Information-Linguistic Model** (ILM), aimed at the "soft understanding" of texts, **mostly by linguistic methods of analysis**. The task of ILM is to bridge the gap between Information Systems and purely Linguistic Systems. Starting from the second one (= understanding any "textual" material) and using all available information resources, ILM aims at building the more meaningful units than those achievable in both approaches separately.

The principle task of semantic analysis (SemAn) in ILM is constructing a dynamic structure that ensures transition from linguistic units (mostly lexemes, the nodes of a syntactic or syntactic-semantic structure) to complex content units of different kinds, familiar to users. In our model it is a primary SemR (Semantic Space, SemSpace), built by using two instruments: the initial Grammar of Semantic Relations (SemRel), that is, syntagmatic and paradigmatic organization of the metalanguage of SemRels, and a Semantic Dictionary (SemDict), both having to be adapted to different information tasks. The SemSpace as the first level textual structure is to be followed by more than one level of SemRs: SitR, EventR, Textual-Fact-Representation (TF-R), all of them aimed at receiving the conceptual status.

2 Outline of Semantic Grammar

A language for the level of SemAn in our model has as its elementary unit a record of the form R(A,B), with R being a binary semantic relation, and A,B, its terms. The same semantic language (sometimes enriched by specific domain relations) can be used for the representation of any domain-knowledge text. The whole list has about 60 SemRels, but only half of them are used in the course of SemAn. A fragment of **semantic relations** list with textual examples follows:

Author (A,B): *novel* (B) *by Hemingway* (A); *Pushkins* (A) *poem* (B)
Addressee (A,B): *to send* (B) smth. *to the press* (A)
Actant (A,B): general name for any type of actant
Quantity (A,B): *two portions* (A) *of soup* (B)
Aspect (A,B): *classify things* (B) *by weight* (A)
Stage (A,B): *at the beginning* (A) *of the meeting* (B)
Sphere (A,B): *work* (B) *in agriculture* (A)
Condition (A,B): *welding* (B) *at high temperature* (A)
Goal (A,B): *NN arrived* (B) *in N.Y. to discuss* (A) *some linguistic problems*
Similar_to (A,B): *The girl* (A) *resembles her mother* (B); this SemRel appears at the second stage of SemAn by the rule of correction of 2 primary formulas: **1-actant** (A,*resemble*) & **2-actant** (B,*resemble*). Words of the category Rel (in the dictionary, see below) such as *resemble*, *similar*, *equal* etc. behave the same way.

Another part of semantic grammar consists of semantic features (SemF), examples:

ARTEFACT (man-made object): *table*, *weapon*, *sputnik*
SUBSTANCE: *sand*, *clay*, *uranium*
PERCEPTION: *hear*, *see*, *observe*
HARM (smth. connected with life hazards): *war*, *threat*, *damage*; *kill*
INTERVAL: (in time) *week*, *intermission*; (in space) *distance*
INFORMATION: *message*, *advertisement*, *novel*
SPHERE: *industry*, *electronics*, *agriculture*
QUANTUM (a unit of smth): *bit*, *lexeme*
CAUSATION (often used in combination with other characteristics): *launch*, *transform*

The Semantic Grammar of Rels is defined by two axes:

- syntagmatic, that is, rules of combining words with given SemF into formulas. Examples:

Reason (A,B), where SemF(A) = SIT or a whole semantic formula; the same for SemF(B). In the phrase *It happened a cause de Peter* two incomplete Rels will be built: Reason (Sit?, happen) & Actant (Peter, Sit?).

Name (A,B), where SemF(A) = Name (A, ?), SemF(B) = Person or Thing or Device;

- paradigmatic, that is, rules of substitution for terms and relations. Examples:

Identificator (A,B) can be specified as Name (A,B) or Symbol (A,B)

Time(A,B) can be specified by two formulas: Starting point(?,B) and Final point(?,B)

There exist more complex superrelations between semantic units (relations, formulas and other units).

As for SemFeatures, they are also organized in a partial hierarchy; an example:

SITUATION > PREDICATE > PROCESS > ACTION

Thus, if a word has SemF = Action, it may be filled in some formula where SemF Sit is predicted.

The list of SemRels is rather stable, taking into account that many general Rels may be specified by index; for instance, SemRel Loc(A,B) may be specified as Loc-inside, Loc-outside, Loc-under, etc., others are to be approved and specified by the rules of semantic analysis. At the moment, Semantic Grammar is used when describing meanings of words in Russian Semantic Dictionary (especially the valency structure) and in building local interpretations for optional and separated groups or parceled propositions in the course of SemAn. Some subtle properties of the Grammar may be used for compression and other transformations of the SemR. For lack of space, I have limited myself to this short exposition of our experimental semantic metalanguage. The full description would require the formal definition of all ILM entities. My task is to outline the whole model of humanlike understanding of a text and to give the general idea of how it may be implemented in the system with developed semantic component.

3 SemDict RUSLAN as a Tool for Textual Semantic Analysis with Compression

The Russian general semantic dictionary (RUSLAN) is the main instrument for intra-, inter-sentence and whole text interpretation of the initial text (Leontyeva, 1995). RUSLAN was implemented in a DB form (Sokirko, 2001). It contains the information on semantic categories and taxonomic characteristics of each lexeme (Semënova, 2000), its semantic valencies, typical contexts, thesaurus relatives, derivational capabilities, collocations, English translations, domain of usage and some other data (about 50 fields organized in 10 zones). Lexicographic descriptions are entered into the DB RUSLAN using our slightly formalized metalanguage of SemRels.

RUSLAN was designed, first of all, for automatic semantic analysis accompanied by gathering informative semantic nodes (SITs and the like units and their constituents) for a given collection of Russian documents. Recognizing and extracting data on people and institutions (on the basis of SemSpace) and building the corresponding semantic nodes were the initial steps along this path. Units of SemSpace being matched successfully against the given DB of VIP-persons or of Organizations may be considered as Concepts. Many of them were restored from textual abbreviated forms to full names of persons, their functions and institutions taken from special DBs. This system was a kind of IE-system, but based on the syntactic and the local semantic analysis. The partial results looked as a naïve Knowbase.

Some data stored in RUSLAN (e.g., data about syntactic realization of the word's valencies) were used only at the pre-semantic stages of NLP. On the other hand, some fields (both linguistic and informational) of a dictionary entry would provide a sort of tuning to the user's request: you may indicate the degree of compression (of the complete SIT-unit predicted in the Dictionary) that you wish. This procedure proposed for RUSLAN-based semantic analysis permits a user to gather an individual BTF according to his/her interests. I will now dwell on some dictionary fields (traditional and new), open to discussion.

3.1 Some Dictionary Fields

• CAT stands for "category". We single out five categories of lexical units (LU) at the moment. Only three of them are described in RUSLAN and have been involved in primary SemAn, their categories depend on the role/place of a unit in semantic formula R(A,B) and further in the semantic graph. These three main classes are: units that may turn 1. into nodes (A or B), 2. into relations (R), or 3. into a node plus relation R(A,?). Lexemes with "empty" meaning form the forth class: (half)empty words occupying places A or B are accompanied by the relation "REF(?,A)", where the question mark requires to restore the omitted meaningful referent. The fifth class includes words whose meaning is an operation on the already existing part of the SemGraph (*accordingly, so, though, nevertheless,* etc.).

LUs of the first category are the most meaningful units (CAT = NODE). Many NODEs may be specified by subcategories: NODE-Object, NODE-Sit, NODE-Attr. Only these lexemes are supplied with semantic features (SemF) and with information WEIGHT (from 1 to 5): WEIGHT(Sit) = 4 or 5, WEIGHT(Attr) = 3, etc.

• SemF are semantic features (ex. "thing", "animate", "process", "intellectual", "bad", etc.; lexical functions – LF – are used in this field as well). Ex.: ENTRY = *bank*; SemF = ORG, FINance. ENTRY = *okazyvat'*; its own SemF = LF Oper.

• VAL is the set of semantic valencies of the word C; candidates for filling in the slots of valencies are introduced by symbols Ai (i = 1, 2, ..., 7). The notation in VAL field is as follows: R(Ai,C) or R(C,Ai). The notation R,Ai,C; R,C,Ai is also possible. Ex.: ENTRY = *message*; VAL = Agent,A1,C; Addressee,A2,C; Topic,A3,C; Content,A4,C.

Each term of a valency is described separately in an abbreviated form: "SemF1" means "SemF of A1", "SemF2" means "SemF of A2", etc.; the same is true for grammatical characteristics: Gram1, Gram2, etc.

• ADD refers to additional semantic relations among the actants. Ex.: ENTRY = *compensation*; VAL = Agent,A1,C; Addressee,A2,C; Cause,A3,C; Value,A4,C; ADD = 1. Patient,A2,A3; 2. Belongs-to,A4,A2. Ex.: *compensation to NN* (A2) *for the damage* (A3); the value (A4) of compensation belongs to NN (A2).

These important data are to be included in SemSpace; they occur usually in text continuation and may be used for disambiguation and for proving the coherence of the text if it will develop that idea (ex., about A2 and event A3 that happens with NN), for inferences.

• CORR is the set of rules of correctness of the valency structure written in the following form: initial SemRel, ⇒ (symbol of transition), resulting SemRel / if … Ex.: ENTRY = *ruin*; VAL = Agent, A1, C; Ob, A2,C. CORR = Agent, A1,C, =>, Cause, A1,C / SemF1 = non-animate; e.g. *the flood* (A1) *has ruined the village* (A2). The *flood* will be Cause instead of Agent of a situation 'ruin' by this rule: Cause,*flood,ruin*; Ob,*village,ruin.*

• RESTOR is the set of rules for reconstructing a member of the valency structure, in particular, the agent of an action being expressed by infinitive.

In the SIT zone, the most important fields are:

• SIT: the fullest linguistic description of a lexeme of the SIT category. Usually, it comprises all Ai from the VAL-field, ex. SIT = {A1–A5} if there were 5 actants. But one may add other members (from ADD or SIT, or PRECED fields).

• ESit, or ES: the description of elementary situations in the form of a set of semantic relations R,A,B; e.g. ENTRY = *export*; VAL = Agent,A1,C; Ob,A2,C; End-Point,A3,C; ESit = 1. Belongs-to,A2,A1 / SemF1 = "organization"; 2. Loc,A1,A2 / SemF1 = "space", "state"; 3. Belongs,A2,A3; 4. Loc,A3,A2. Further on, elementary situations are referred to as ES1, ES2, etc.

• PRECED: an elementary situation preceding the main Sit (C); PRECED = ES1 or ES2.

• POST: an elementary situation following main Sit, e.g. ENTRY = *export*; POST = ES3 or ES4.

In the PRAGM zone, the important fields are:

• DOMAIN: polit., or econ., or war, or usual.

• WEIGHT: initial semantic importance of the lexeme (specified by linguist). Ex.: *war* 5, *start* 4; *nice* 3, etc.;

• EVENT: event (main situation denoted by the word C and/or one of its actants with the greatest informational weight that may be a nucleus of some event in the indicated domain), e.g. EVENT = A3 (actant A3 of C is announced to be the center of TF to be built); Ex.: ENTRY = *to help*; VAL = Agent,A1,C; Addressee,A2,C; Content,A3,C; Reason,A4,C; EVENT = A3.

• INFER: a standard inference in the form of a production rule: If SIT1, then SIT2; If SIT2, then SIT5.

• PRESUP: presupposition (the name of a situation already introduced in a field or formulated by the linguist, that is indispensable for C to be true).

• EVAL: evaluation ("+", "-", "0", or "?", where "?" signifies that the evaluation depends on certain conditions); e.g. ENTRY = *export*; EVAL = ? / (A2) (evaluation of C is inherited from A2): EVAL (*to export drugs*) = "-"(bad). This SemF is used mostly when analyzing criminal texts.

• LOG: a more complex situation that characterizes the logic of the event and is to be formulated in terms of SITs and production rules. (For ex., to connect notions *crime – trial – sentence – punishment*).

I will dwell on two fields of RUSLAN which differ from the ECD approach. These are VAL and GRAM fields, which are filled in the "top-bottom" way. You write in VAL all SemRels expected to be

expressed in the text, not splitting them into strong and optional relations&fillers. The GRAM field consists of two parts (written as SYN: MORPH). The SYN part predicts by what syntactic class or syntactic role this actant in adopted system of SynR can be expressed. The MORPH part lists all morphological means used to express this role. Each VAL may have more than one superficial expression. This gives the rules of SemAn more flexibility.

Another big part of the RUSLAN dictionary is the dictionary of **words-relations** (punctuation marks, conjunctions, prepositions, etc.). The WEIGHT of Rels depends on the WEIGHT of their terms. Really, this second category dictionary is an extension of the SemGrammar. The words of the third category are **words-parameters**, which occupy the first place in R(A,B) formula, being semantically dependent on B and thus having the lesser weight; the name of R is a generic notion for them: Param(***height***, *man*); Part(***arm***, *human*). Many words-parameters repeat the R-name. Ex.: Time(***time***, SIT); Part(***part***, *body*). The WEIGHT of them is usually 3. For the lack of space I can't develop these themes.

3.2 Dictionary Implementations and the State of Affairs

The first version (ROSS) of our semantic dictionary has been an important component of semantic analysis in the FRAP MT-system (French-to-Russian Automatic Translation) in the National Translation Centre. Then the Russian-to-English version was developed and has been included in the text understanding system POLITEXT elaborated in the Academy Institute of USA and Canada (ISCRAN). Some modifications were made to include a part of that dictionary (Russian and English parts separately) into a Russian-to-English MT-system (Sokirko, 2001). The Russian part of the system was used as the main tool for semantic interpretation of texts in the Internet network version (www.aot.ru). The Bulgarian Academy of Sciences began working on a Bulgarian simplified version of the dictionary. RUSLAN-1 is being implemented and upgraded at Moscow State University; at the moment, it contains about 12, 000 semantic entries. It may be called by any application dealing with syntactic, semantic and pragmatic analysis of Russian texts; it has thus the status of reusable dictionary resource. (However, for the last two years, the work on development of the SemDict was practically stopped.)

4 Types/Stages of Understanding in ILM

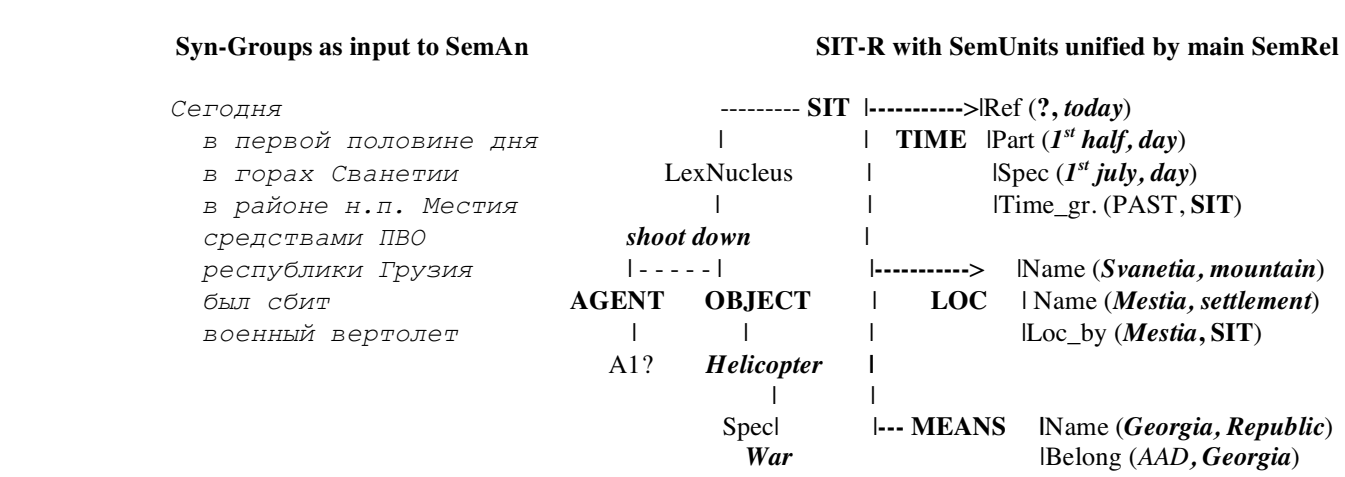
Stages of understanding, the corresponding representations and the instruments of analysis (roughly, dictionaries) adopted in ILM and partially realized in some systems (see Leontyeva, 1987 & later issues) are shown in the following table:

Types of understanding	Stages of analysis	Representations (R)	Dictionaries
Local understanding (within one sentence)	GraphAn, MorphAn, SynAn, primary SemAn	GraphR, LexR, MorphR, SynR, local SemRs of separate sentences	Grammatical Dicts, Special linguistic Dicts, ROSS/ RUSLAN
Global understanding (inter-sentence analysis)	Proper SemAn and SitAn (of utterance, text)	Semantic Space, SitR, Global Sem(SIT)R	RUSLAN, ling. frames, SIT-Schemata
Relative (user-oriented or domain-oriented) analysis/ understanding	Interpretation of SemR in terms of domain and user request / interest	Special R-s, Textual Facts (TF)-R, Event-R, TF-Base or BTF	Thesaurus, databases, Special Frames (of Orgs, Persons, Geogr., etc.)
Understanding in terms of another language	Translation, Text Generation Systems	English variants of Sit-Rs, Event-R, TF-R, BTF	Russ-Eng SemDicts

As for the fourth stage of understanding, it may be the translation of BTF in English or in another language; we hope that it is possible to use for this purpose an already existing TG-system.

Let us consider a simple illustration of three levels of Representation (SynR, SitR, TF-R) of a short newspaper text:

1 июля. Сегодня в первой половине дня в горах Сванетии в районе населенного пункта Местия средствами ПВО республики Грузия был сбит военный вертолет.
'1st july. | Today | in the first half of the day | in Svanetia mountains | near the settlement of Mestia | by means of AAD | of the Republic of Georgia | was shot down | a military helicopter.'



The TF-representation of the sentence under consideration is as follows:

EVENT = ***Сбили военный вертолет*** (1,2,3) – The node built according to the rule of meaningful SIT-name
1. **AGENT** = ***ПВО Грузии*** (textual SemNode deduced as the Agent due to absence of other candidates)
2. **TIME** = ***1989 г.- 1 июля*** (the node restored from the data of the text, and 1989 – from the publication date of the newspaper)
3. **LOC** = ***Грузия - Сванетия - Местия*** (the full name restored from the Geographical DB).

We have obtained a TF-representation of this message manually, by using Sem-Rules, the SIT-Structure, the external DB (Geography), the rule of inference (node A1 may be also generalized as Georgia), and Text attributes (see the full TIME-node).

5 Semantic Space

The local SemAn begins with sentence-for-sentence semantic interpretation (in terms of the adopted metalanguage) of the initial text. The basic mechanism of local SemAn has been implemented as a procedure of translating syntactic formulas r(a,b) of the SynR – adopted as input structure – into semantic formulas R(A,B):

1. attr (two,books) ⇒ Quantity (TWO,BOOK)
2. indirect obj (by Stern,books) ⇒ Author(STERN,BOOK) or more exactly:
   Author(STERN,1.), that is, Stern is the author of BOTH BOOKs (1. refers to the 1st formula).

Thus, at the beginning of SemAn we obtain rather a SynSemR, which corresponds to the first level of text understanding.

The local SemAn deals with the valencies of each lexeme. Those valencies that are not saturated within the limits of one sentence or do not satisfy some rules of semantic grammar and dictionaries are transmitted into SemR as invalid formulas: R(A,?), or R(?,B). These gaps in SemR become the important signs of text incoherence. Sometimes they may be filled in by reference to those portions of specific knowledge that are fixed in the SemDict or they have to be restored at the next stage – the global SemAn.

SemSpace may have conflicting formulas due to splitting of meanings of a single syntactic node; in this case, they are combined by the superrelation "INCOMPATIBLE(C,D)". Ex. *Сопротивление* 'resistance' *проводника* 'of conductor' will give two formulas: 1. Agent(*conductor* as "human", *resistance* as "action") and 2. Parameter(*resistance* as "parameter", *conductor* as "device"); they will be two members of the superrelation "INCOMPATIBLE(1.,2.)" in the same SemSpace – not to be split into different variants of SemR. Similar conflicts may be resolved by further steps of SemAn.

Now I will mention some other particularities of SemSpace that differ from standard linguistic SemRs in a principled way.

From the structural point of view, SemSpace is a sequence of formulas having simple syntax of binary relations R(A,B), where the second member is normally the main one and has therefore the greater

weight. (I will omit the procedure of semantic correction of some formulas if this main member has "empty" meaning in dictionary, as well as some other phenomena concerning our metalanguage formal demands.) As for members of SemRels, they are mostly lexemes, but can also be formulas, grammatical elements or even empty slots. Some SemRels may link not only lexical units, but nonterminal symbols as well. Ex. Repres (SIT5,Prop2), which means 'SIT5 is computed as the main representative of the Proposition 2'; Compos(25 Prop,Txt) 'Text consists of 25 propositions', etc. Those compositional relations are used in SemAn for references, when moving through the text, comparing formulas and making inferences, ex. Equal(Sit2 of Prop1,Sit7 of Prop8).

The divisions between sentences may be ignored. SemSpace without boundaries becomes a **continuous** textual structure being a basis for proper Semantic Analysis. As such, it has useful properties. SemSpace is a **flexible** structure: you may throw out or add some formulas, you may mix and change the order of sentences, utterances etc., you may analyze SemSpace from the last to the first sentence if it necessary for a particular task, etc. For example, the signature under a long order would permit you to reconstruct later the referent of the omitted name of the Author-valency of the first textual lexeme *PRIKAZYVAJU* 'I order': *1., 2.,* etc. As for the Content-valency of the same lexeme, it must be specified as the set of paragraphs introduced by numbers: *1., 2., 3.* etc. To collect such information you need to take into account even the visual form (GraphR) of the document under analysis.

As an experimental structure, SemSpace has a **dynamic character**: it allows the system to build meaningful complex units, to restore locally omitted units and to introduce lost "weak" groups in the proper slots throughout the text. Alike or comparable formulas may be found in the textual continuous structure due to homogeneous syntax and "natural" meanings of SemRel names. Moreover, the new compressed complex units of the next level (SitR, EventR) enter the same SemSpace along with its primary lexical units, etc.

SemSpace is an **adaptive structure**: its units may be compared with any external parts of the knowledge given by the user. If the user's request has been analyzed using the same metalanguage, it may be possible to replace step by step some text units by analogous terms of user's interest. The result will be an individual information representation, in the form of a frame corresponding to the ordered part of knowledge. As for the SemDict, the transition from one given Domain to another one does not lead to changing the dictionary. It will concern only few fields that are to be "tuned" to the Domain.

SemSpace as a kind of linguistic SemR adds one more dimension to the whole text understanding (WTU) – a **vague** one. SemSpace can include incomplete formulas, conflicting and redundant data, heterogeneous terms of SemRels, unknown words and other units without information.

6 Situational Analysis & Representation

The main problem of ILM is to develop mechanisms of global SemAn. The deficiencies of SemSpace have a **creative** character: they serve as indicators for restoring omitted parts of meaning, for compressing the redundant pieces of SemR, etc. All these processes aim at gathering greater units, thus compressing several empty formulas of the primary SemSpace into one complete unit. Global SemR is to be constructed according to the rules of semantic grammar of the following form: R1(A,?) & R2(?,B) ⇒ R3(A,B) // under certain conditions. Another rule deletes formulas that duplicate each other in full or partially. One more resource for constructing bigger units from elementary Sits is the standard structure of textual linguistic SIT (the part of it may be seen in the example above). Complex SITs can be constructed outside the limits of a sentence as well as inside a sentence.

I would like to emphasize the idea that the coherent text analysis is an analysis through generation of complex units, this generation beginning inside analysis. The processes of analysis and generation within text overlap much more tightly than one might think. Several elementary Sits may be combined into a full SIT according to the canonical linguistic SIT-structure being applied to every simple sentence of a text. The most effective is gathering of isolated, incomplete and parceled sentences. Ex. from A. Blok: «*Ночь. Улица. Фонарь. Аптека*» will be analyzed at the SIT level as follows: *Ночь* 'night' is the TIME of some SIT that is absent. *Улица* 'street' is LOC of some SIT that is absent. *Фонарь* 'lantern' is an object of type 'device' that may give light (LF from Dict). *Аптека* 'pharmacy' is some place (LOC) of some SIT that is absent, it is bigger than 'lantern', but 'street' is bigger than 'pharmacy' (from pragmatic knowledge introduced manually). So we may deduce that all those entities must be parts

of the same SIT that is absent (no any other alternative). But we can bind them to this unknown SIT (?SIT) using their own semantic characteristics that predict the SemRels. The whole description will be as such: Time (*Ночь*,?SIT); Loc (*Улица*,?SIT);. Obj (*Фонарь*,?SIT='action'); Loc (*Аптека*,?SIT); Loc (*Улица*,*Аптека*); Ref (?,SIT). The last formula permits to look for the referent of the unknown SIT further in this text (it is a created valency of this complex utterance under analysis). Redundant units in the global SemSpace may be eliminated, using rich dictionary resources of modern lexical semantics schools (ECD, see Mel'čuk, 1984-1999 and Apresjan *et al*., 2000, not to mention the number of special Thesauri). But this way needs further study.

If we collect all SITs being built in the course of analysis of a given text, this will be the SIT-Representation that adds one more structure to the set of compressed products. Each SIT must be accompanied by its own frame, which can be empty to a certain extent (any absolutely empty SIT will be deleted from SemR). The main SIT (if it is meaningful) with its subordinate Sits may turn to be a textual fact (TF) representing the principal content of the text under analysis (if no TF has been built). Other variants of summarization may be possible by agreement (only SITs, SIT + EVENT, SIT + some important lexemes etc.). As for EVENT-unit it may be built as linguistic unit on the base of dictionary prediction (see field EVENT) or by the procedure of generalization of several SITs.

7 Relative Analysis and TF-Representation

Primary SemSpace is a linguistic structure indifferent to a notional sphere. All transformations applicable to SemSpace aim at constructing meaningful units – SITs and TFs. If built from linguistic material, they remain purely linguistic entities. Meanwhile the ultimate task of **specialized/professional texts analysis** is receiving a KnowBase consisting of **conceptual entities**. How to examine lexical units for conceptual status? The simplest way to prove that a linguistic unit (LU) may be transformed into conceptual entity, or concept (CNC), is to match the LU, be it simple or complex, against some existing Domain sources introduced as **counter-text**. The results of comparison being successful (according to certain conditions), the LU may be declared a conceptual entity (LU ⇒ CNC). In this way we could obtain (theoretically) one more kind of SemR consisting mostly of concepts, notions, etc. It may be called the naïve Knowledge Base (KnowBase), in our case the Base of Textual Facts (BTF). It may be a common BTF or individual KnowBases for different specialists. Each of them forms a new dimension of text content.

The specialized text SemAn is a type of Pragmatic analysis (PragmAn). Such SemAn has been implemented (in ISCRAN), matching syntactic noun phrases against the Thesaurus of Russian political life. The results obtained were far from ideal: syntactic and terminological units boundaries were often in contradiction.

The following experiment consisted in matching semantic groups from SynSemR against lists of domain objects such as Names of politicians, Geographical notions, Institutions, and Political functions. These domain objects have to receive their own syntactic or semantic information before the matching procedure, so that the process was "top-down" one. Those groups (noun phrases) that succeeded in matching received the status of denotation and a greater information weight and may enter the final DBs.

The information about denotational status of textual units is very important when procedures of further global SemAn proceed within the limits of Semantic Space. The completeness of DBs taken as counter-texts will lead to the success of PragmAn and further BTF-construction. If there are no matches (the PragmAn, or domain Analysis, fails, has zero results), the initiative is passed to proper LinguAn.

Formally, a "Textual Fact" is a multi-term predicate where the terms are maximally meaningful notions (i.e., they have maximum informational weight for the given text and/or for the given domain and pragmatic orientation). It's worth mentioning that a TF as a multi-term predicate is not the same as that of syntactic structure: the terms of the former has to be gathered across the whole text – usually not according to syntactically "strong" actants of an appropriate word-predicate. Moreover, many literary texts (being analyzed without "counter-text") will have as their TF the name of a simple object, not of a predicate. Ex.: in I. Bunin's novel "The Gentleman from San-Francisco," the title turns to be a TF, in our metalanguage – "Start_point(*S.-F.*, *Gentleman*)", and as its terms would be listed the words = "actions" of that Gentleman. I see here the similarity with some DBs where the name of some place, or animal, bird etc. is announced as the main predicate of a table, and all properties (can fly, etc.) of such an entity form fillers in DB fields.

The TF-structure in linguistics-based systems is to be built by summarizing all SitRs constructed in conformity with rules of global semantic analysis and with properties of the coherent text structure. As for the practical tasks, TF and its terms are to be computed taking into account pragmatic orientation – what Events you are looking for. Below is one of 6 TFs built manually on the Corpus of about 100 short texts from the newspaper "Obshchaja gazeta" in 1991. It was the first attempt to formulate some rules of TF-analysis (introduced in a man-machine scenario), the terms were taken from the verbal material.

TF = COUP D'ETAT (1,2,3,4,5)
Variant: *seizure of power*
1. Agent = *the USSR State Emergency Committee (SEC)*
   Variant = t*he Soviet leadership*
   Identification = *G.Yanayev, V.Pavlov, O.Baklanov, B.Pugo, V.Starodubtsev, A.Tizyakov, V.Kryuchkov, D.Yazov*
2. Counter-agent = [*former power, President Gorbachev*]
3. Cause = *destabilization of political and economic situation in the USSR*
4. Goal = *to overcome economic and political crisis in the USSR*
5. Time = *from August 19, 1991*

It is obvious from this description that it was a "relative fact," true only for the texts in question, for the given data, etc. To enter the historical DB of real FACTs, our **Textual Fact** as a linguistic unit must be compared with many TFs from other reliable textual sources. I regard this task as a serious challenge to linguists.

The final representation of a text is to be supplied by the text-description with its own title (or number), source (author) of information, composition, etc. Actually, it will be the frame of the document. As for the content, each document may be represented by Situations (SIT), Events (EVENT) and/or Textual Facts (TF), each accompanied by its own frame, temporal data, names of places of events etc. So, the BTF for a given corpus of texts has to be a condensed structure, a result of comparison and generation of new units.

8 Conclusion

Based on the belief that a complete computable understanding of a whole text is impossible to achieve, I conceive the **text compressing ability** as an obligatory property of NLP systems. We have tested our SemAn and SemDict mostly on real documents (Decrees of Russian President, criminal reports, newspaper messages and titles, etc.). SemSpace was the first whole text structure being implemented, and this experimental structure merits to be regarded attentively. An uncertainty of different kinds – redundancy, incompleteness, ambiguity, and contradiction to the SemGrammar rules – appears in this initial SemR in an explicit form. **The task of eliminating text "defects" and that of text compression meet each other** in the design of ILM, thus creating the moving force for the next stage – the proper SemAn. The phenomenon of the **multiplicity of possible views** on the same thing, be it text or political life, made me abandon the intention of building one unique well-formed structure of a text. Maybe such semantic structure as our "handicapped" SemSpace will ensure a new solution for many NLP problems.

Many formal definitions of separate word meanings seem not corresponding to the real usage in the huge realm of texts, general and specific, scientific and metaphoric, rather vague than exact, etc. (cf. Altman & Polguere, 2003 and Wanner, 2003, who discussed the difficulties of programming the ECD definitions). Being too formal, they involve great difficulties in text automatic treatment by NLP systems. We have proposed the **more "soft" dictionary descriptions** of words, expressions, symbols etc., using SemRel language. As a subset of NL it is a semi-formal one, but it has many advantages of NL "naturality." Our SemRel language was tested on difficult examples, rich in linguistic violations and distortions, which no purely linguistic theory would be able to manage. Meanwhile the "invalid" SemSpace built in the course of SemAn leaves the possibility to use the whole semantic context. The choice of that or another variant of lexeme meaning may be based on common sense rules such as: *One person can't be in two places at the same time*, or *Agent and its action are not to be separated* (*at the same time and space*), and the like. This way would relieve an applied system of many bottleneck problems in the proper linguistic works.

The more cardinal idea consists in automatic analysis of classical dictionary descriptions in terms of our SemRels. By building SemSpace for Dictionary entries we may obtain natural and powerful resources of lexico-semantic knowledge. Using this information as additional counter-text would advance the problem of textual SemAn. But this task as a whole needs some new technique (of analysis and matching).

The **counter-text** idea seems to make the task of **including Special Knowledge** in NLP not so unreal. We admit that Knowledge may be involved in the NLP system by small portions **in a usual textual form** that can be formulated by the User (his request, list of desired terms, etc.) – to examine the idea. To involve the formal descriptions of different Domains in NLP-system may constitute the next, not immediate task (see Nirenburg, 2004). But I suppose it will be more realistic to face the domain problem as a kind of machine translation problem using the same metalanguage for initial text analysis as well as for domain text analysis followed by comparison (and adaptation) of those "foreign-like" languages.

Acknowledgements

Two linguistic phenomena that have led me out of limits of SynR to the whole text problems are syntactic ellipse and semantics of prepositions in the field of Machine Translation. Both topics were launched by Igor Mel'čuk, the supervisor of my Ph.D. dissertation, 50 years ago. I am very grateful as well to Lidija Iordanskaja for her important remarks on this text, and I appreciate her interest in this work.

References

Altman, Joel & Alain Polguère. 2003. *La BDef: base de definitions derivée du Dictionnaire explicative et combinatoire //* MTT 2003, Paris, 16-18 juin 2003, 43-54

Apresjan, Ju, *et al*. 2000. *Новый объяснительный словарь синонимов русского языка.* Jazyki slavjanskix kul´tur, Moskva

Grishman, Ralph. 1999. *Information extraction: Techniques and Challenges*. Internet

Leontyeva, Nina.1987. Stages of Information Analysis of Natural Language Texts. *Int. Forum Inf. and Docum*. 12: 4, 8-14

Leontyeva, Nina. 1995. ROSS: Semantic Dictionary for Text Understanding and Summarization. *META*, 40: 1, 43-54

Mani, Inderjeet. 2001. Automatic Summarization. In R. Mitkov, ed., *Natural Language Processing*, vol. 3. Benjamins, Amsterdam/Philadelphia, 286 p.

McKeown K. *Text Generation*. Cambridge, 1988.

Mel'čuk, Igor *et al*. 1984 - 1999. *Dictionnaire explicative et combinatoire du francais contemporain. Recherches lexico-semantiques. Vol. I - IV*. Les Presses de l'Université de Montréal, Montréal

MUC - Message Understanding Conferences, 1-7.

Nirenburg, Sergey & Victor Raskin. 2004. *Ontological Semantics*. MIT Press, Cambridge, MA

Semënova, Sofia. 2000. Word Taxonomy Fields of one Russian General-Purpose Semantic Dictionary: Descriptor Selection, Analysis of Representation Possibilities. In *DIALOG-2000*, vol. 2: 308-316

Sokirko, Alexej. 2001. *Semantic Dictionaries in Automatic Processing of Texts*. Ph.D. dissertation. Moskva

Wanner, Leo. 2003. Definitions of Lexical Meaning: Some Reflections on Purpose and Structure. In *MTT 2003*, Paris, 55-65

Separating Different Lexemes of The Korean Adjective I-

Geun-Seok Lim
Postdoctoral researcher at the University of Montreal, OLST
geun-seok.lim@umontreal.ca

Abstract

This paper's goal is to separate and describe the lexicographic senses (= lexemes) of the Korean adjective I-. The previous studies of I- are limited to several typical uses of I-, while we believe that we should examine the whole range of I- uses so that all of its lexemes can be examined. The paper proposes typological, semantic and syntactic considerations that can be used to separate the lexemes. Through the application of these considerations and the discussion of the characteristics of I- , this paper shows that I- can have 19 lexemes.

1 Introduction

This paper aims to separate and describe the lexemes of the Korean adjective I-[1] without using a specific methodology of a theory.[2] There have been a lot of papers related to I- but they were limited only to its typical uses. As we will see, some typical uses can't cover the usage completely. The previous studies were biased toward two important questions[3]: 1) whether I- is a word or a suffix and 2) on the assumption that I- is a word, what is its part of speech (a verb, an adjective or a new type of part of speech?). But no approach can reach the right conclusion without the sufficient considerations of considering the whole range uses. And we believe that if we consider the whole range of uses of I-, we will not call I- COPULA, SUFFIX or SEOSULGYEOK JOSA ("a case marker of predicate"), which are terms many Korean grammarians want to call I-. To clarify the identity of a linguistic unit, we should show all of its uses and expose the relations among them.

Let's take a look at various examples of uses of I-.[4]

(1) i saram-i Kim Cheol-Su-(i-)[5]-da.
this person-SUBJ Kim Cheol-Su-(«I-»)-DECL
'this person is Cheol-Su Kim'

(2) geu-neun haksaeng-i-da.
he-THEME student-«I-»-DECL
'he is a student'

(3) jangmi-neun ppalgansaek-i-da.
rose-THEME red.color-«I-»-DECL
'the rose is red-color'

---

[1] Usually the Korean adjective I- is translated as BE in English. But in this paper it will not be called COPULA because it has various lexemes which cannot be described and covered within the concept of copula.

[2] But the basic philosophy of this paper concerning the meaning of a linguistic unit is the same as in Mel'čuk *et al.* (1995).

[3] Because of the limitation of pages, representative papers for each issue were not mentioned in this paper. Someone interesting the development of I- argument can refer to K.I. Nam (2004) and H.P. Im (2006).

[4] The abbreviations are as follows: COM: commutative ending, COORD: coordinate ending, COP: copula, DAT: dative ending, DECL: declarative ending, EXCL : exclamation, GER: gerund ending, «I-»: the Korean adjective I-, MOD: modifying ending, OBJ: objective ending, POST: postposition, PRES: present (prefinal ending), SUBJ: subjective ending, THEME: thematic ending

[5] When I- comes after a vowel, it is usually truncated (or disappears) in colloquial Korean.

(4) i-hoisa gisul-eun gukjejeok-i-da.
this-company technology-THEME international-«I-»-DECL
'this company's technology is international (or quite good)'
(5) Cheol-Su-ga jebeop-i-da.
Cheol-Su-THEME quite-«I-»-DECL
'Cheol-Su is quite good at his job/study/etc'
(6) jeongchi-pan-i gagwan-i-da.
politics-situation-THEME spectacle-«I-»-DECL
'(lit.) the situation of politics is spectacular / (implication) The situation of politics is disgusting'
(7) geu-neun yojeum buleo gongbu-e yeolsim-i-da
he-THEME nowadays French language study-DAT hard-«I-»-DECL
'nowadays he is hard in studying French'
(8) na-neun geu-gyeoljeong-e bandae-(i-)da
I-THEME that-conclusion-DAT objection-(«I-»-)DECL
'I object to that conclusion / Me, there is objection to that conclusion'
(9) Cheol-Su-neun Yeong-Hui-wa nammae-(i-)da.
Cheol-Su-THEME Yeong-Hui-COM brother and sister-(«I-»-)DECL
'Cheol-Su and Yeong-Hui are brother and sister'
(10) na-neun jigeum jip-i-da.
I-THEME now home-«I-»-DECL
'now I am at home.'
(11) Oneul-eun samwol o-il-i-da.
today-THEME March 5-day-«I-»-DECL
'today is March 5th,'
(12) ppang-eul meok-eun geos-eun Cheol-Su-(i-)da.
bread-OBJ eat-MOD entity-MOD Cheol-Su-(«I-»-)DECL
'it is Cheol-Su who ate the bread'
(13) na-neun Jajang-i-da.
I-THEME Jajang-«I-»-DECL
'(lit.) I am Jajang(Korean food) /(implication) I prefer Jajang, or I'll eat/prepare/cook/etc Jajang'
(14) san-i ontong nun-i-da.
mountain-SUBJ entirely snow-«I-»-DECL
'the mountain is entirely covered by snow'
(15) wa! jjang-i-da.
EXCL chief/very much-«I-»-DECL
'wow, it's great'
(16) wa! nun-i-da.
EXCL snow-«I-»-DECL
'wow, there is snow'
(17) nu-eoseo Tteok-meok-gi-(i-)da.
lie-CONN Tteok-eat-GER-(«I-»-)DECL
'(lit.) [It] is Eating Tteok(Korean food) in the position of lying /(implication) It's so easy'
(18) na-neun bap-eul meok-neun jung-i-da.
I-THEME rice-OBJ eat-MOD middle-»I-»-DECL
'I am eating rice'
(19) na-neun jip-e doraga-l geos-i-da.
I-THEME house-DAT go back-MOD thing-«I-»-DECL
'I will go back home'

The main examples mentioned in the various papers on 'I-' are (1)-(3). The others were ignored; although some papers dealt with some of (4)-(15), it was done sporadically and never explained the relation

between them. 'i-'s in (4)-(9) can be said that they have the same meaning with (3), or 'property', if only the lexical meaning is considered somehow. But if we examine the grammatical meaning and the characteristics of the syntactic behavior, we will get to know they should be distinguished. 'i-' in (10) has the meaning 'to localize a thing in a place' and 'i-' in (11) is the meaning 'to localize a thing in a time'. (12) is the example of cleft sentence. 'i-' in (13) has an intangible meaning but we can say it has the meaning, roughly speaking, 'to prefer a thing'. 'i-' in (14) has the meaning 'to be covered with'. 'i-'s in (15)-(17) have only one argument and have different meanings. 'i-'s in (18) and (19) are used in the constructions of grammatical collocation and grammatical idiom.[6] Anyway distinguishing the lexemes of I- should be done before the explanation of the relation.

2 Some considerations for lexeme separation

In order to distinguish the lexemes of I-, we will use three types of considerations ─ typological, semantic and syntactical. These three types of considerations were chosen because 1) the lexemes of the Korean adjective I- are found in other languages, 2) I- has various lexical and grammatical meanings and 3) the difference between lexemes, we believe, should be verified by the difference in syntactic behavior.

2.1 Typological considerations

There are languages that have more than one copula. If, when a typological consideration dividing the copulas is applied to the Korean I- it makes any semantic or syntactic difference of the examples, then the dividing consideration should be regarded as an important one. And the dividing criteria should be proved by syntactic considerations.

A detailed typological description of the copula is given in Pustet (2003). Some of the examples she mentioned are chosen for our discussion of I-; which will help us to discriminate between lexemes of I-.

[1] identificational *vs*. ascriptive: é *vs*. hécha in Lakota
The criterion that differentiates between these two lexemes[7] is that of uniqueness vs. non-uniqueness of extra-linguistic referents of predicate phrases in the universe of discourse.
[2] semantic class *vs*. property concept: kh i *vs*. pen in Thai
In Thai the nouns combining with different copulas have different meanings: (20).

(20) a. nî: kh i : să:mlìa:m
this COPa triangle
'this is a triangle'
b. (kracok') nî: pen să:mlìa:m
mirror this COPb triangle
'this (mirror) is triangular'

[3] nominal *vs*. adjectival *vs*. verbal *vs*. quantificational, temporal and participial : ye…ye, ka, b ɛ , dɔ̀n in Bambara
Different copulas are chosen according to the part of speech of the lexical unit which combines with it. See Bambara's examples in (21).

(21) a. nìn ye námása ye
this COPa banana COPa
'this is a banana'

[6] grammatical collocation and grammatical idiom are my own term to refer to specific Korean constructions which look like a collocation and an idiom, respectively.
[7] Poustet(2003:29) used the word "type" but in our context it is a "lexeme".

b. so ka sùrun
house COPb small
'house is small'
c. ne b ɛ taa
1SB COPc leave
'I am leaving'
d. caman dɔ̀n
many COPd
'there are many'

[4] normal nominal *vs.* nominal of place and time : ɛ́ *vs.* ɑ́ in Nuer
In some languages different copulas combine with the different types of nominal. Especially place and time nominals have their own copula.

Pustet introduces further examples of multi-copulas. But only preceding examples are useful to divide lexemes of I-. From her distinctions we can use 4 considerations; we will check out 1) whether the meaning of I- is identificational or ascriptive, 2) class or property if the meaning is a ascriptive, 3) the part of speech of a lexical unit combining with I-, 4) the type of nominal. See 3.1 for the results of applying the typological approach to I-.

2.2 Semantic considerations

As mentioned above, most papers which dealt with I- have focused on mainly (1)-(3). So it is quite natural many grammarians consider I- as a empty unit which doesn't have any specific lexical meaning. But if we want to account for all the lexemes of I-, we should not say that I- is a semantically empty unit. Because many lexemes have lexical meanings. In this section we will use three semantic considerations for distinguishing the lexemes of I-.

[1] lexical meaning *vs.* grammatical meaning
First of all, we should check out whether the examples of I- have lexical meanings or just grammatical meanings. But the distinction between lexical and grammatical meaning is not strictly dichotomous as many other linguistic phenomena. Furthermore, the judgment of lexical meaning can be different depending on the different theoretical bases researchers stand on. So here we will use the following consideration: If a lexeme of I- is more concrete in comparison with other lexemes, then we will consider it as having a lexical meaning.
[2] semantic actants
The number and constraints on semantic actants of I- are various. Most grammarians normally take it for granted that I- has two semantic actants and that the characteristics of the actants of each lexeme are quite similar or even the same. But we have counterexamples which will be shown in 3.2.
[3] substitutability
Finally, we can replace 'i-'s with other words based on our intuition. In this paper we will substitute I- only with synonyms for the sake of simplicity but theoretically synonymous expressions can be tested for this purpose. Anyway the closer the meanings of 'i-'s are, the more synonymous expressions or synonyms they will share.

2.3 Syntactic considerations

We have discussed the typological and semantic considerations for dividing the lexemes of I-. But there are still several loopholes.[8] In order to make up those weak points and expose the differences of each lexeme, we will use two main considerations. First one is that if a specific syntactic characteristic occurs if

[8] First, typological consideration has already connected to semantic and syntactic considerations. So we need to develop the typological consideration into more detailed syntactic consideration. Second, it is very difficult to describe the difference of each lexeme's meaning when the meanings are very close without using a sophisticated meta-language.

and only if the combination between preceding element and the specific lexeme 'i-' is given, we can regard the use as one lexeme. Second is that all the typological and semantic differences must be attested parallel by different syntactic behavior.[9]

We will test eight features to expose the particularities of each lexeme.

[1] dropping of I-
The possibility of dropping is quite different for different lexemes of I-. We admit that we should use statistic data extracted from spoken corpora. But in this paper we will just follow our intuition. So we will give "3" points when the probability of dropping I- is high; "2" when average, "1" when low, "0" if "impossible".
[2] switching between X and Y
If a lexeme of I- has two arguments(X and Y), then we will switch the position of the two arguments. Some lexemes allow for a switch, others do not but the meaning changes.
[3] adding a modifier: from 'X Y-COP' to 'Y-COP-MODIFIER X'
If a lexeme of I- has two arguments, we will transform a predicate construction into a modifying one. Some lexemes do not allow for this conversion and some do but the meaning changes. If the modification is possible, we will write "Yes", otherwise "No".
[4] restriction on combinations with TAM(tense-aspect-modality endings)
The restriction in combination with TAM is quite different for differing lexemes of I-. We will give "3" points when the restriction in combining with TAM is high; "2" when average, "1" when low, "0" when "no restriction".
[5] restriction in inserting the honorific prefinal-ending -SI-
The possibility of inserting the honorific ending -SI- is different for differing lexemes of I-. We will give "3" points when the restriction in inserting the honorific prefinal-ending -SI- is high; "2" when average, "1" when low, "0" when "no restriction".
[6] arranging Ys in a row by using a coordinate ending -GO
We will test the expandability of the argument Y. If possible we will assign "Yes", otherwise "No".
[7] negation
We will try to convert the examples into negative sentences and fill up the check list of each lexeme with the negative forms; different lexemes of I- react differently to this test.
[8] modification by an adjective or an adverb

We will check whether the construction composed of I- and the preceding linguistic unit can be modified by an adjective or an adverb. And if necessary, we will write the modifying words in the check list.

3 Applying considerations

3.1 Applying typological considerations

Firstly, let's look at 'identificational' vs. 'ascriptive'. Among the examples of (1)-(19), only (1) and (12) can be described as 'identificational'; most examples are 'ascriptive' and some cannot be divided by this consideration. The Y of (1) can be identified by hearer without ambiguity. But 'i-' within the cleft sentence like (12) cannot always be combined with identificational Y. See (22), 'bae-ga gopa-seo' cannot be referred to as (1).

(22) geu-ga bab-eul meok-eun geos-eun bae-ga gopa-seo-(i-)da.
he-SUBJ rice-OBJ eat-MOD thing-THEM belly-SUBJ hungry-CONN-(«I-»-)DECL
'the reason why he ate the rice is that he was hungry'

[9] The reason why we did not mention morphology is that the grammatical differences between the lexemes of I- have nothing to do with the morphology. For instance, the conjugation of tense should be dealt with in syntax, not in morphology. But it is not within the main discussion topic in this paper, so the further discussion will be skipped.

Furthermore the semantic and syntactic properties of (1) and (12) are quite different. Here we can have two categories but they should be subdivided. For instance, the types of 'ascriptive' are too heterogeneous to maintain one category.

Secondly, the distinction between 'class' and 'property' is quite useful. 'haksaeng' of (2) is the name of a class and 'ppalgansaek' of (3) is the name of a property. The difference of the two types is that the former has a possibility for dichotomy but the latter has a degree of the property. This semantic difference makes a syntactic difference. But this distinction between 'class' and 'property' is not always clear. The syntactic difference of the two lexemes and difficulty of distinction between 'class' and 'property' will be shown in 3.3.

Thirdly, in terms of typological perspective, the parts of speech of Ys are various: nominal, adjectival, verbal, quantificational, temporal and participial. In Korean, normally a noun can be put in the position of Y but an adjective and a verb cannot. But there are some lexemes which can admits to combine with other elements, not only a noun. This will be mentioned in 3.3.

Fourthly, we can use the division of "normal nominal" vs. "nominal of place and time". But before applying the consideration to the Korean I-, we need to check the relation of inclusion; that is, "normal nominal" can include "nominal of place and time". See (1) and (24).

(1) i saram-i Kim Cheol-Su-(i-)[10]-da.
this person-SUBJ Kim Cheol-Su-(«I-»)-DECL
'this person is Cheol-Su Kim'

(23) yeogi-ga Seoul-i-da.
here-SUBJ Seoul-«I-»-DECL
'here is Seoul'

The argument Y of (23) is a noun of place but the lexeme of I- of (23) is the same as (1), not the same as (10). If we test the semantic characteristics and syntactic behavior of I- of (23), we come to know that it is exactly the same as (1).

3.2 Applying semantic considerations

As mentioned in 2.2, we will investigate the meaning of I-. Firstly, the most typical grammatical ones, of the 19 uses, are (1) and (2). Because the meaning is the most general and it simply connects the two arguments without adding a specific meaning. On the contrary, the examples of (13) and (14) add a more specific meaning. For instance, 'i-' of (14) has the meaning 'be full of / be covered with'.

Secondly, we will examine the number of actants and the characteristics of them. Some lexemes of I- have two actants, some just inherit actants from the linguistic unit which combines with I-, some have only one actant. The explanation of I- having two actants will be skipped. Let's look at (8). At first glance, I- of (8) seems to have 3 actants. But this mistaken observation can be corrected by the fact that BANDAE of (8) has two semantic actants as we can confirm in (24). So we can say that I- of (8) has inherited the two syntactic actants from the predicate noun BANDAE. In this paper we will consider 'i-' of (8) as a support verb.[11]

(8) na-neun geu-gyeoljeong-e bandae-(i-)da
I-THEME that-conclusion-DAT objection-(«I-»-)DECL
'I object to that conclusion / Me, there is objection to that conclusion'

---

[10] When I- comes after a vowel, it is usually truncated (or disappeared) in colloquial Korean.

[11] In completing this paper, Prof. Mel'cuk gave me a great deal of advices. It was much appreciated. He told me 'i-' of (8) look similar to (16), or might even have the same meaning. I admit the lexical meaning of I- of (8) is quite similar to (16), but the grammatical meaning is different with (16). But this topic is not important to this paper. Someday I hope to have an opportunity to discuss further about the Korean support verb I-.

(24) geu-gyeoljeong-e daeha-n[12] na-eui bandae.
that-conclusion-DAT to face-Mod na-GEN objection
'my objection to the conclusion'

The most interesting example of I- is (16) which has only one actant. In that sentence the speaker just says 'there is snow'. If we give two actants to I- of (16) and change (16) as (25), then the meaning will be different.

(16) wa! nun-i-da.
EXCL snow-«I-»-DECL
'wow, there is snow'
(25) igeos-i nun-i-da.
this-SUBJ snow-«I-»-DECL
'this is snow'

Thirdly, we will replace I- with other words more or less synonymous with I- in the given context. We will give each word its own number to distinguish it; 1 for MAT-('be right'), 2 for DONGILHA-('be identified'), 3 for HA-('do'), 4 for ITT-('exist'), 5 for DOI-('become').

The result of applying typological and semantic consideration to I- will be shown in following table.

(26) The result of applying typological and semantic consideration

	1-1	1-2	1-3	1-4	2-1	2-2	2-3
1	Id	△	Noun	Noun	Gra	2	1, 2
2	As	Class	Noun	Noun	Gra	2	1, 5
3	As	Property	Noun	Noun	Gra	2	1, 5
4	As	Property	Modifying noun	Modifying noun	Gra	Inherit	X
5	As	Property	Adverb	Adverb	Gra	Inherit	X
6	As	Property	Unique morpheme	Unique morpheme	Gra	Inherit	1
7	As	Property	Root	Root	Lex	Inherit	X
8	As	Property	Predicate noun	Predicate noun	Lex	Inherit	1, 5
9	As	Property	Relation Noun	Relation noun	Lex	Inherit	1
10	△	△	Noun	Noun	Lex	2	4
11	△	△	Noun	Noun	Lex	2	1, 4
12	(Id)	△	Noun	Noun	Gra	2	1, 2?
13	△	△	Noun	Noun	Lex	2	many words[13]
14	△	△	Noun	Noun	Lex	2	4
15	As	Property	Adverb	Adverb	Lex	1	X
16	△	Normal N	Noun	Noun	Lex	1	4
17	As	Property	Noun phrase	Noun phrase	Lex	1	X
18	△	△	(Bound) Noun	(Bound) noun	Gra	X	X
19	△	△	Bound Noun	Bound noun	Gra	X	X

[12] "e daeha-n" is a multiple-word which has grammatical meaning 'about'.
[13] I- of (13) can be substituted by many words according to the context.

3.3 Applying syntactic considerations

As mentioned in 2.3, some syntactic characteristics just show in specific combinations between I- and the preceding element of I-. I- of (7) is combined with a root of a word without an affix. The preceding element, or YEOLSIM-, can be a syntactic unit only when it combines with I- as (7). But it can not be used without the suffix –HI when it does not combine with –I (see 27).

(7) geu-neun yojeum buleo gongbu-e yeolsim-i-da
he-THEME nowadays French language study-DAT hard-«I-»-DECL
'nowadays he is hard in studying French'

(27) geu-neun gongbu-reul *yeolsim/yeolsim-hi ha-da
he-THEME study-OBJ hard do-DECL
'he studies hard French'

We cannot show the whole process of applying 2.2 in detail because of the page limit. So we will show several applications of syntactic consideration.

First, we will check the probability of I-'s omission.[14] Some lexemes allow for the drop of I- but some never do: see (28) where I- can not be dropped. (28) comes from (4).

(28) *I hoisa gisul-eun gukjejeok.
this company technology-THEME international

And then we will switch the two arguments if I- has two semantic actants. Some lexemes allow the switch but the meaning changes.[15] (29) is the switched sentence from (2). The sentence of (29) is acceptable but the meaning changes. I- of (29) seems quite similar to that of (1); the difference is in the communicative structure (Theme/Rheme inversion).

(29) haksaeng-i geu-(i-)da.[16]
student-SUBJ he-(«I-»-)-DECL
'the student is he'

We also examine the restriction of honorific prefinal-ending -SI-. Some examples, for example (16), have the restriction of combining with -SI-. 'i-' of (16) is usually used when a speaker is not expecting the present of someone or something in that place and in that time but suddenly the person or the thing appears. So basically 'i-' of (16) can combine with any respectable noun. But as we see in (30) where we used respectable noun SEONSAENGNIM('teacher'), it is not acceptable.

(30) ??wa! seonsaengnim-i-si-da.[17]
EXCL teacher-«I-»-SI-DECL
'wow, there is (our) teacher'

---

[14] In this paper we cannot discuss at length the omission but we can take a short look at the different meanings when I- is omitted. See (1') which comes from (1). We haven't mentioned the ambiguity of (1) so far, but actually (1) can be interpreted as 'identificational' or 'ascriptive'; usually 'identification', rarely 'ascriptive'. But when I- is omitted as (1'), the meaning of the sentence is only as 'identificational'. So we need to be careful when dealing with the omission.

(1) i saram-i Kim Cheol-Su-(i-)-da.
this person-SUBJ Kim Cheol-Su-(«I-»)-DECL
'this person is Cheol-Su Kim' / 'this person's name is Cheol-Su Kim.'

(1') i saram-i Kim Cheol-Su
this person-SUBJ Kim Cheol-Su
'this person is Cheol-Su Kim'

[15] "Pos(m.c.)" in the check list stands for the case, or when the switch is possible but the meaning changes.

[16] This is not a completely acceptable sentence but still can be understood.

Finally the modifications of an adjective and an adverb are also dealt with. Some constructions which combine with I- can be modified only by an adjective, some by an adverb, and some by both an adjective and an adverb. For instance, (31)-(33) can be modified by adjectives. But the modification adverbs are quite different. The predicate construction of (31), which comes from (2), cannot be modified by adverbs, whereas the predicate constructions of (32) and (33), which come from (1) and (3), can be modified by adverbs. (32) and (33) still take different adverbs for their modifying unit.

(31) geu-neun *maeu/*baro haksaeng-i-da.
he-THEME very/just student-«I-»-DECL
'he is a very/just student'
(32) i saram-i *maeu/baro Kim CheolSu-(i-)da.
this person-SUBJ very/just Kim CheolSu-(«I-»-)DECL
'this is the person, Cheol-Su Kim'
(33) jangmi-neun maeu/*baro ppalgansaek-i-da.
rose-THEME very/just red color-«I-»-DECL
'(lit.) The rose is very red-color'

The reason why (31) cannot have adverb-modifying, and (32) and (33) have different adverbs is that the semantic sorts of nouns of (31)-(33) which combine with I- are different. HAKSAENG of (31) refers to a class, Kim Cheol-Su of (32) refers to a specific person, and PPALGANSAEK of (33) denotes a property, respectively. But this distinction is not always clear-cut. For example, the noun BUJA is usually considered to denote a class. So we will guess that (34) is similar to (31). But as we see in (35), the construction "buja-(i)-da" can be modified by the adverb "maeu". So we can say that the I- of (34) can be interpreted as the lexeme of (3) as well as the lexeme of (2).

(34) geu-neun buja-(i-)da
he-THEME rich person-«I-»-DECL
'he is a rich person'
(35) geu-neun maeu buja-(i-)da
he-THEME very rich person-«I-»-DECL
'(lit) he is a very rich person'

It may be said that the modifying of MAEU is relevant just to the semantic type of the modified noun, or BUJA. But we should not overlook the fact that without the help of I-, BUJA cannot be modified by any adverb. See (36) and (37).

(36) *geunyeo-neun maeu buja namja-reul johaha-n-da
she-THEME very rich man-OBJ like-PRES-DECL
*'she likes very rich man'
(37) geunyeo-neun maeu buja-i-n namja-reul johaha-n-da
she-THEME very rich-«I-»-MOD man-OBJ like-PRES-DECL
*'she likes the man who is very rich'

(38) The result of applying typological and syntactic considerations.

	3-1	3-2	3-3	3-4	3-5	3-6	3-7	3-8
1	3	Pos	Pos(m.c.)	0	0	Yes	ani-	Adj, adv
2	2	Pos(m.c.)	Pos	0	0	Yes	ani-	Adj

[17] But if people remove the exclamation WA and use this sentence to express the meaning like (2) or (3), then it can be an acceptable utterance.

3	2	Pos(m.c.)	Pos	0	0	Yes	ani-	Adj, adv
4	0	n/a	n/a	0	0	Yes	-ji anh-	Adv
5	0	n/a	n/a	0	1	Yes	Impos	Adv
6	1	n/a	n/a	0	1	Yes	Impos	Adv
7	0	n/a	n/a	0	0	Yes	-ji anh-	Adv
8	1	n/a	n/a	0	0	Yes	ani-	Adj, adv
9	1	n/a	n/a	0	0	Yes	ani-	Adj
10	2	Pos(m.c.)	Pos	2	0	No	ani-	Adj
11	2	Pos(m.c.)	Pos	0	2	Yes	ani-	Adj
12	1	Impos	Impos	0	0	Yes	ani-	Adj, adv
13	3	Pos(m.c.)	Pos(m.c.)	2	1	No	ani-	Adj
14	1	Impos	Impos	1	3	No	Impos	Adj, adv
15	2	n/a	n/a	0	2	No	Impos	Adv
16	2	n/a	n/a	3	3	No	Impos	Adj
17	2	n/a	n/a	0	3	Yes	ani-	Adv
18	1	n/a	n/a	1	1	n/a	ani-	n/a
19	0	n/a	n/a	1	3	n/a	ani-	n/a

4 Conclusions

At this point we have examined the whole set of examples which were given in the beginning of this paper. We conclude that every 'i-'s in these examples can be described as a separate lexicographic lexeme. It is possible for some of the lexemes to be united for particular purpose but this is a topic for the next study: first, we have to distinguish all lexemes with the utmost accuracy.

References

Nam, Kil Im. 2004, *hyeondae gukeo ida gumun yeon-go* (The study of the *ida* construction in Modern Korean), Hanguk munhwasa.

Mok, Jeong Su. 2006, *ida-reul gineungdongsaro bunseokhaeya haneun myeotgaji iyu* (Some arguments for defending 'ida' as a support verb, *Eomunyeongu* 136, Eomunyeonguhoi.

Im, Hong Pin. 2006, *jeongche balkimeui hyeongyongsa ida munjewa yeone* (The problem of identifying adjective ida and collocation), *Eonmunhak yeongueui neolbiwa gipi*, Yeokrak.

Lim, Geun-Seok. 2006, *Hangukeo yeoneo yeongu* (A study on Korean collocation), Ph. D. Dissertation, Seoul National University.

Lim, Dong Hoon, idagumuneui jesimunjeok seonggyeok (The Theticity of Korean *Ida* Constructions), *gokeohak* 45, gukeohakhoi.

Haspelmath, Martin. et al. (eds) 2005, The World Atlas of Language Structures, Oxford University Press.

Mel'čuk, Igor A., Clas, André & Polguère, Alain. 1995, Introduction à la lexicologie explicative et combinatoire, Louvain-la-Neuve: Duculot.

Pustet, Regina. 2003, *Copulas: Universals in the Categorization of the Lexicon*, Oxford University Press.

Paddy Fields: A Topological Description of Chinese Word Order[1]

Pierre Magistry
CLCLP, TIGP Academia Sinica
Graduate Institute of Linguistics
National Taiwan University
pierre@magistry.fr

Kim Gerdes
LPP (CNRS), Signes (INRIA)
ILPGA
Sorbonne Nouvelle
kim.gerdes@univ-paris3.fr

Abstract

This paper shows how Mandarin Chinese word order can be described in topological terms. After discussing the difficulties in using syntactic dependency and basic word order phenomena of Chinese, we provide the foundations of a formal topological grammar that links a syntactic dependency tree to all of the possible corresponding word orders. We present the formal rules that allow the generation of simple sentences, as well as the more complex *ba* and *bei*-constructions and serial verb constructions.

1 Introduction

This paper is an account of the work in progress regarding a topological description of some basic word order phenomena of Mandarin Chinese. The topological model (Gerdes Kahane, 2001) is a powerful model for linearization. In the Meaning-Text Theory, it can model the interface between surface syntax and the next level, traditionally called morphological representation, that we refer to as the "topological level", because we not only construct word order and prosodic breaks of different sizes, but a fully-fledged constituent tree. This tree can then provide the basis for the computation of the prosodic groups and pauses. The topological model is a formalization of the traditional analysis of German (Drach 1937) and has been shown to allow for an elegant description of word order phenomena like scrambling in V2 or verb final languages that mix syntactic and communicative constraints. The basic idea is that the sentence is constructed from fixed places, also called positions or fields, which have different constraints on the number of their occupants. All word order constraints are described in this manner: If a word has to precede another word, we don't use relative placement rules, but these words are placed into different fields. If their mutual order is free (under the given communicative information), they go into the same field.

At first sight, it may seem like an overkill to apply this model to a language like Chinese, sometimes described as having a very restricted word order because of its limited morphology. Alternatively, Her 2003 describe "inversion constructions" in LFG with "simplified Lexical Mapping Theory". In this approach, a change in argument order is considered as a lexical operation, putting the full burden on a multitude of lexical entries. However, following (LaPolla 95), we will show that Chinese has a fairly complex word order, depending mainly on communicative constraints. Placing this work in the Meaning-Text Framework allows us to have unordered dependency trees at the syntactic level; the linearization process can then be described in topological terms. All the rules and examples we provide have been implemented and tested with the DepLin software (http://gerdes.fr/soft/deplin/), which assures that obtrusive interaction between different rules or surplus word orders have not been overseen. The rules have, however, not been tested in parsing, although this is feasible (for example by transcoding the rules in Lexical Functional Grammar (LFG), Clément et al. 2002). Aside from the difficulty involved in

[1] The authors, not being Chinese natives, are deeply indebted to the helpful comments and innumerable grammaticality evaluations by Hsieh Shu-kai, Liu Yeh-hsin and Jun Miao. We have also benefited greatly from Sylvain Kahane's and three anonymous reviewers' comments on our work. Any shortcomings remaining after help from these colleagues are, of course, entirely our own responsibility.

working on written text when we want to include many oral word orders in our account, we would encounter the word separation and ambiguity problem that most rule-based approaches face when parsing Chinese, often obscuring the underlying analysis of word order phenomena (cf. for example the importance that segmentation takes in the development of the Chinese Lexical Functional Grammar in Fang&King 2007). We believe that this is another example of the usefulness of the prevalence of the synthesis direction in the Meaning-Text Theory. It allows concentrating on the non-coincidental properties of language, while keeping in mind the bidirectional character of the rules provided.

2 Adequateness of the Meaning-Text Model

It is nonetheless an important question to ask whether the Meaning-Text Model provides an appropriate framework for a language like Chinese. The pipeline model with semantic, deep syntactic, and surface syntactic representations needs to be discussed for a language where the usual difference between the semantic representation and the surface syntactic representation does not apply easily: Semantemes become full form words. In Chinese, not only do we lack morphologically-based differences between different categories, for example between nouns and verbs, they commonly keep the same valency in whatever syntactic position the words appear. In the following sentences, *ai*, just as its English translation 'love', appears as verb or as a head-noun without any morphological change.

(1)	你 愛 她	/	你 對 她 的 愛 有 多少
	nǐ ài tā		nǐ duì tā de ài yǒu duōshao
	you love she		you to her DE love ~have many-few
	'You love her.'		*'How deep is your love for her?'*

Contrary to English, however, where we have morphological tests (changing person, time, and number) for a clear distinction between the two categories, in Chinese, the only observable difference between the two 'ai' is the syntactic context, for example the appearance of DE, a genitive particle, when *ai* could be called a noun phrase. More generally, the semantic-syntax interface remains the role to provide function words, appearing on the syntactic level (like *de, ba,* and *bei* presented in section XXX4) and to choose pronouns (or, more often, the absence of pronouns) when realizing predicates.

Yet, the main reason for stipulating doubt on the appropriateness of MTT is the central position this model gives to dependency, including the prominent place of syntactic functions. Although categorical borders may be very different in Chinese (see for example Huang 1997), to our knowledge, nobody doubts the existence of categories as a whole. Things are different with syntactic functions: LaPolla 1993 convincingly shows that the usual criteria for subjecthood or objecthood do not exist in Chinese and argues in favor of a completely semantic and pragmatic analysis of the language, meaning that semantic roles such as agent and beneficiary, coupled with communicative values like topic and focus, are sufficient to describe word order constraints in Chinese. At this time, we cannot discuss whether Chinese has truly grammaticalized the subject role, and it is possible that the term "agent", even in the surface dependency, would be more appropriate. However, we remain with the usual functional terms *subject* and *object* whenever we have the syntactic realization of an agent in a dependency tree. We will nevertheless use semantically tainted terms like *goal* if a common equivalent for the syntactic relation cannot be found among the usual syntactic functions.

In this approach, we follow the common practice in computational and formal description of Chinese such as the work on a Chinese LFG in the Palo Alto Research Center (Fang&King 2007) or the work of Haitao Liu 2007 on syntactic dependency structures for Chinese, using the "European" terms as function names wherever possible. His work on a Chinese dependency treebank has demonstrated that the dependency approach can give important insights into the structure of the Chinese language.

3 Simple Structures and first formalization

We start our description with a simple dependency structure with a transitive verb:

(2) 我 昨天 買 了 書

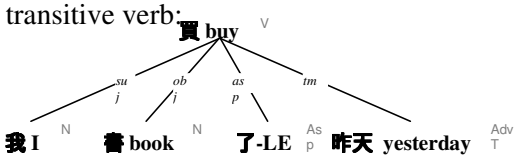


Figure 1: Simple dependency tree

Wǒ zuótiān mǎi le shū
I yesterday buy ASP book
I bought books yesterday.

Note that we have the two arguments, the subject and the object, realized as a pronoun and a bare noun. Temporal and spatial relations behave slightly differently than other modifiers and we have to introduce a specific modifier relation, *loc*, which hints further at the close connection between semantic and syntactic relations in Chinese. The aspectual marker LE, marking the accomplishment[2] will be treated it as a separate word with a special function: *asp*.

3.1 Topicalization, word order possibilities, and communicative structure

Chinese is said to be an SVO language, which may be misleading considering that S and O functions are potentially irrelevant. LaPolla 95 suggests that that Chinese should be described as "verb medial" language where "Topical or non-focal NPs occur preverbally and focal and or non-topical NPs occur post-verbally." The typical order given in **Erreur ! Source du renvoi introuvable.** is in fact the most communicatively neutral, corresponding to Li & Thompson 81 (chapter 4.1.3 D) "sentences with no topic", i.e. it can constitute an answer to the thetic question: *What is going on?*. Topicalization of various dependents of the verb is possible with different communicative structures. Multiple topicalization is possible, too, in particular in spoken language. This can lead to very different word orders for the same dependency structure.

(3)

這 本 書 我 買 了	/	昨天 書 我 買 了
Zhè běn shū wǒ mǎi le		zuótiān shū wǒ mǎi le
this Classifier(Cl) book I buy ASP		yesterday books I buy ASP
This book, I bought (it).		*Yesterday, books, I bought (some).*

Li & Thompson describe this possibility for *shu* (book) to be in the topic position as in (3)**Erreur ! Source du renvoi introuvable.**. They also remark that the topic position cannot be occupied by indefinite NP and that the interpretation of bare nouns is constrained to be either definite or generic. Interestingly enough, we should add that when in object (post-verbal) position, a bare noun is either generic or indefinite but cannot be interpreted as definite. These differences of possible interpretations seem to be closely related to the communicative value born by the bare noun.

(4)

我 買 了 書	/	書 我 買 了
wǒ mǎi le shū		shū wǒ mǎi le
I buy ASP book		book I buy ASP
'I bought (a) book(s)'		*'The book , I bought it'* or *'Books, I bought'*
**'I bought the book'*		** 'A book, I bought'* or ** I bought a book.'*

Note that our analysis differs slightly from Li&Thompson's presentation of Chinese simple declarative sentences. We allow the subject to be placed in a topical position, creating a different constituent structure, whereas Li&Thompson talk about "Sentences in Which the Subject and the Topic are Identical" vs. "Sentences with no subject". The difference lies in the definition of the subject position, the one they give making it impossible to distinguish those two positions when the topic is an agent. We consider, although we cannot show this here, that the communicative difference also appears prosodically, and we capture this kind of (spoken) word order possibility by allowing more than one element in topic position.

The aspectual marker *le* occupies a position in close proximity to the verb, from which it can only be separated by a verbal resultative (in so called Verb-Resultative compounds) or a specific kind of object in Verb-Object compounds, which are collocational or idiomatic and thus lexically constrained. (5)(5) is an example of a Verb-Object Compound where the bare noun *fàn* can appear before the aspect marker, but it

[2] It is generally agreed upon the fact that Chinese has two different markers LE, the other type is called "Current relevant state" (CRS) which always has to go in the last position in the sentence. This place is the last position of our micro domain. LE is sometimes designated as a verbal suffix or as an auxiliary, the lack of segmenting characters making those two explanations plausible.

can also be topicalized or appear after *le* as in (6)(6)**Erreur ! Source du renvoi introuvable.**. In the latter cases, *fan* could also have dependents that would specify the meal. This is not possible when *fan* occupies the position between chi and le where it can only appear as a bare noun.

(5) 我吃飯了

Wǒ chīfàn le
I eat meal ASP
I ate.

(6) 我吃了飯 / 飯我吃了

wǒ chī le fàn	fàn wǒ chī le
I eat ASP meal	meal I ate ASP
I ate.	*I've eaten* (more like *"lunch, I already had"*).

For the dependency tree presented above, the topicalization possibilities amount to 8 different word orders (of the 120 theoretically possible orders). They correspond to 16 different communicative structures, which reflect different possibilities for the intonation structure in spoken language.

3.2 Domains and placement rules

Topological grammars can include communicative constraints directly in the rules. In this work, however, we provide a grammar that gives all the possible word orders, independently of the communicative partition, but it is straightforward to specialize the proposed rules with communicative restrictions. The terms we use for the description of these possibilities stem from the syntactic description of oral French (Blanche-Benveniste 1990) where we distinguish the "macrosyntactic" domain providing places for all extraction and topicalization phenomena from a core syntax, called "microsyntax", with the common order constraints and places for all verbal arguments (used when the arguments are rhematic). Moreover, we consider that Chinese verbs provide places for some of its closer dependents. We call this the "verbal domain".

The macrosyntactic domain only has two fields: The thema-field and the main field. Note that this macrosyntactic division in two main fields roughly corresponds to Chao 1968's description of the Chinese clause structure as simply topic and comment. The micro domain distinguishes four places to express the ordering constraints: subject field, verbal field, object field, and SVC field. The verbal domain has the following fields: circ(umstantial) field, ba-bei-field, negative field, verbal field, verbal object field, and the field for the aspectual (marker). We obtain the following *domain descriptions* including the placement constraints for each field:

Macro domain:	macro-d	= Topic* Micro-field
Micro domain:	micro-d	= subject? verbal! object? Svc? CRS?
Verbal domain:	verbal-d	= circ* neg? ba-bei? verb! v-obj? Asp?
Nominal domain (simplified):	nd	= dem? Num? Cl? atr* noun!

In order to provide places for their (direct or indirect) dependents, words will open these domains under certain conditions, given in the *domain creation rules*: We describe the domain creation rules as a tuple (original field, category, communicative value, domain to be created, final field). A dependant word can occupy an existing position under conditions called the *placement rules*: They can depend on the following values: (the governor's category, the governor's communicative value, the governor's field, the syntactic relation between governor and dependant, the dependent's category, the dependent's communicative value, the field where the dependent can go into)[3].

[3] Topological grammars can also control extraposition by a hierarchy of domains permeabilities, not used in this simplified extract of the full grammar.

Below we present the complete rule set needed for the description of the word orders of the examples of section 3. The verb at the root of the dependency tree is placed first and will follow these box creation rules, and the following placement rules may apply to his nominal dependents. (Communicative values would have to be defined at this step):

Initial field	Category	Domain created	Final field
I	V	macro-d	Micro-field
Micro-field	V	micro-d	verbal
verbal	V	verbal-d	verb

Governor POS	Governor's field	relation	Dependent POS	Dependent field	comment
V	verb	subj	N	subject	Neutral subject
V	verb	obj	N	object	Neutral object
V	verb	suj	N	Topic	Topicalized subject
V	verb	obj	N	Object	Topicalized object

The aspect marker LE will be placed by the first rule, the nominal dependents by the following two rules, and the temporal adverbial by the remaining rules:

Governor POS	Governor's field	relation	Dependent POS	Dependent field	comment
V	verb	asp	ASP	Asp	Aspect marker
N	noun	atr	CL	Cl	Classifier in NP
CL	Cl	qc	Num	num	Numeral in NP
V	verb	loc	AdvT	circ	circumstancial
V	verb	loc	AdvT	Topic	Topicalized circumstancial

Then the nominal dependents may open nominal domains in various positions:

Initial field	Category	Domain created	Final field
Subject	N	nd	noun
Object	N	nd	noun
Topic	N	nd	noun

These rules suffice to compute the different linear structures for the above dependency tree. We show here the topological structure for a simple topicalization of “book”, corresponding to sentence 3 below:

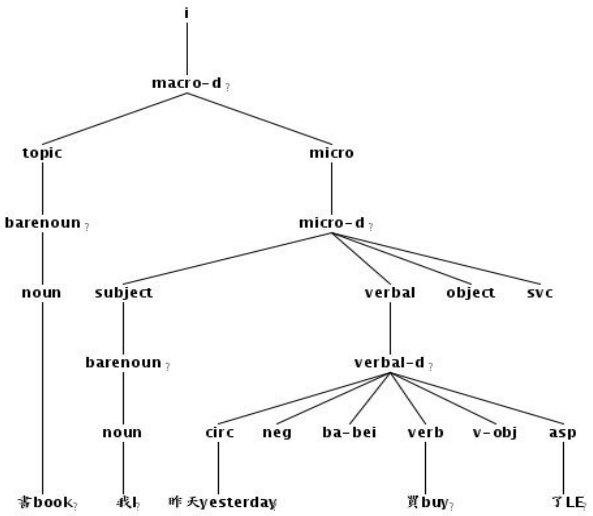


Figure 2: A topological tree = structure 3 below

The complete list of all the analyses is given below. Note that we can have different topological structures for the same word order. This means that an analysis of written text using these rules will find for one sentence up to three topological trees, corresponding to the same dependency tree (a *topological ambiguity*). In the generation approach going all the way to sound output, however, these structures are essential for the computation of the prosodic structures.

15. i[*macro-d* topic[*nd* noun我I] topic昨天yesterday topic[*nd* noun書book] micro[*micro-d* verbal[*verbal-d* verb買buy asp了LE]]]

Number of structures with the same word order: 1

---

1. i[*macro-d* micro[*micro-d* subject[*nd* noun我I] verbal[*verbal-d* circ昨天yesterday verb買buy asp了LE] object[*nd* noun書book]]]

6. i[*macro-d* topic[*nd* noun我I] micro[*micro-d* verbal[*verbal-d* circ昨天yesterday verb買buy asp了LE] object[*nd* noun書book]]]

8. i[*macro-d* topic[*nd* noun我I] topic昨天yesterday micro[*micro-d* verbal[*verbal-d* verb買buy asp了LE] object[*nd* noun書book]]]

Number of structures with the same word order: 3

---

13. i[macro-d topic[nd noun我I] topic[nd noun書book] micro[micro-d verbal[verbal-d circ昨天yesterday verb買buy asp了LE]]]

16. i[macro-d topic[nd noun我I] topic[nd noun書book] topic昨天yesterday micro[micro-d verbal[verbal-d verb買buy asp了LE]]]

Number of structures with the same word order: 2

---

14. i[macro-d topic昨天yesterday topic[nd noun我I] topic[nd noun書book] micro[micro-d verbal[verbal-d verb買buy asp了LE]]]

Number of structures with the same word order: 1

---

2. i[macro-d topic昨天yesterday micro[micro-d subject[nd noun我I] verbal[verbal-d verb買buy asp了LE] object[nd noun書book]]]

7. i[macro-d topic昨天yesterday topic[nd noun我I] micro[micro-d verbal[verbal-d verb買buy asp了LE] object[nd noun書book]]]

Number of structures with the same word order: 2

---

4. i[macro-d topic昨天yesterday topic[nd noun書book] micro[micro-d subject[nd noun我I] verbal[verbal-d verb買buy asp了LE]]]

10. i[macro-d topic昨天yesterday topic[nd noun書book] topic[nd noun我I] micro[micro-d verbal[verbal-d verb買buy asp了LE]]]

Number of structures with the same word order: 2

---

3. i[macro-d topic[nd noun書book] micro[micro-d subject[nd noun我I] verbal[verbal-d circ昨天yesterday verb買buy asp了LE]]]

9. i[macro-d topic[nd noun書book] topic[nd noun我I] micro[micro-d verbal[verbal-d circ昨天yesterday verb買buy asp了LE]]]

12. i[macro-d topic[nd noun書book] topic[nd noun我I] topic昨天yesterday micro[micro-d verbal[verbal-d verb買buy asp了LE]]]

Number of structures with the same word order: 3

---

5. i[macro-d topic[nd noun書book] topic昨天yesterday micro[micro-d subject[nd noun我I] verbal[verbal-d verb買buy asp了LE]]]

11. i[macro-d topic[nd noun書book] topic昨天yesterday topic[nd noun我I] micro[micro-d verbal[verbal-d verb買buy asp了LE]]]

Number of structures with the same word order: 2

4 The “Ba” and “Bei” constructions

The Ba and Bei constructions have been widely discussed in Chinese linguistics literature. The latter is sometimes referred to as “passive”. Topologically, they share the same position (since they are in complementary distribution), between the negation marker and the verb, and they behave like a preposition, opening a position for a NP. **Erreur ! Source du renvoi introuvable.** and (8) provide examples. The difference between these two prepositions is that Ba will take a patient as a complement where Bei will take an agent. As we mentioned in Section 2, we consider them to appear at the syntactic level.

(7) 我 把 那 本 書 買走了

Wǒ bǎ nà běn shū mǎizǒu le
I BA this Cl. Book buy ASP
I bought this book.

(8) 那 本 書 被 我 買走 了

nà běn shū bèi wǒ mǎizǒu le
This Cl. Book BEI I buy ASP
I bought this book / this book was bought by me.

Note that a bare noun would have to be interpreted as definite or generic just like topics. In other words, they cannot introduce new information to the discourse. This confirms the idea that new information has to be postverbal. The position of negation adverbs leads us to locate these constructions inside the verbal domain, just between the verb and the negation adverb:

(9) 我 沒 把 那 本 書 買走了

Wǒ méi bǎ nà běn shū mǎizǒu le
I have-not BA this Cl book buy
‘I did not buy this book.’

(10) *我 把 那 本 書 沒 買走了

Wǒ bǎ nà běn shū méi mǎizǒu le
I BA this Cl book have-not buy

An important point is that Ba and Bei cannot be topicalized, neither can the depending NP:

(11) *書 我 把 買走 了 / *把 書 我 買走 了

shū wǒ bǎ mǎizǒu le / bǎ shū wǒ mǎizǒu le

book I BA buy ASP / BA shu I buy ASP

The position for Ba and Bei is opened by the verb and already included in the rules we have presented in section 3.2. Now we need to define their placement rules and their own domain that will hold the dependent NP. We have two domains: bei-d = bei subject and ba-d = ba object

Domain creation and placement rules:

Initial field	Category	Domain created	Final field
ba-bei	BA	ba-d	ba
ba-bei	BEI	bei-d	bei

Governor POS	Governor's field	relation	Dependent POS	Dependent field
V	verb	pat-obj	BA	ba-bei
V	verb	agt-obj	BEI	ba-bei
BA	ba	comp	N	object
BEI	bei	comp	N	subject

These additions to our grammar suffice to generate the more restricted word orders: With a 6 words tree, only two different word orders are possible, corresponding to 5 different topological trees (for 720 theoretical possibilities) :

1. [*macro-d* micro[*micro-d* subject[*nd* noun我I] verbal[*verbal-d* circ昨天yesterday ba-bei[*ba-d* ba把BA object[*nd* noun書book]] verb買走 buy asp了LE]]]
3. [*macro-d* topic[*nd* noun我I] micro[*micro-d* verbal[*verbal-d* circ昨天yesterday ba-bei[*ba-d* ba把BA object[*nd* noun書book]] verb買走 buy asp了LE]]]
5. [*macro-d* topic[*nd* noun我I] topic昨天yesterday micro[*micro-d* verbal[*verbal-d* ba-bei[*ba-d* ba把BA object[*nd* noun書book]] verb買走 buy asp了LE]]]

Number of structures with the same word order: 3

---

2. [*macro-d* topic昨天yesterday micro[*micro-d* subject[*nd* noun我I] verbal[*verbal-d* ba-bei[*ba-d* ba把BA object[*nd* noun書book]] verb買走 buy asp了LE]]]
4. [*macro-d* topic昨天yesterday topic[*nd* noun我I] micro[*micro-d* verbal[*verbal-d* ba-bei[*ba-d* ba把BA object[*nd* noun書book]] verb買走 buy asp了LE]]]

Number of structures with the same word order: 2

5 Serial Verbs Constructions

Chinese is also known to have Serial Verbs Constructions (SVC) even though it has been convincingly argued that this term in Chinese linguistics subsumes a multitude of different constructions (Paul 2004)[4].

5.1 Determining the direction of dependency: purpose vs. circumstantial and SVC

It would be impossible to cover all the various phenomena subsumed in the term SVC, therefore, in this paper, we will now focus on the first type described in Li&Thompson 1981, in which the SVC expresses two separate but related events. The same surface form (NP V NP V NP) can lead to four different relations between the two verbs. The relation can be either (i) consecutive, (ii) purpose, (iii) alternating, or (iv) circumstance. Where (i) and (iii) are often ambiguous as well as (ii) and (iv). Paul 2004 argues that not only different interpretations are possible but they should be regarded as different constituent structures with the same surface form. And, following Paul, amongst those different structures, only the purpose relation is a proper SVC. He describes (i) and (iii) as a VP coordination (iv) as an adjunct and (iii) as a proper SVC (ambiguous in surface with (ii). Here we recall Paul's analysis of the SVC:

(12) 我們 開會 討論 這個 問題

Wǒmen kāihuì tǎolùn zhège wèntí[5]

[4] Among them are some structures that should not be called SVC because they resemble phenomena very common in various languages including languages without SVC, like sentential subjects. Nevertheless, some so-called SVC in Chinese are comparable with structures of African languages well known for their SVC (But even in this case, a close look to characterize structural differences amongst languages is needed, see Wu 2002, Paul 2004)

[5] We have to note here that when asked about this sentence, some native speakers (of Mandarin spoken in Taiwan) don't even notice the ambiguity (in favor of 17b) or said to have a strong preference for the SVC interpretation.

we hold-meeting discuss this Cl problem

a. Wǒmen [VP [adjunct pro ∅ kāihuì] [VP tǎolùn zhège wèntí]
'We'll discuss that problem holding a meeting'
b. Wǒmen [VP kāihuì [*purpose clause* tǎolùn zhège wèntí]]
'We'll hold a meeting to discuss this problem'

5.2 Syntactic structure and additional topological rules

This case can be addressed easily in terms of MTT at a syntactic level. Since the difference is clearly in the relation between the two verbs, we should define two different syntactic relations, in opposing directions, one for the circumstances and one for the purpose, see the two corresponding dependency trees below. We can then define topological rules to account for different linear groupings. By doing this, we pay attention to various constraints on word order that reflect the structural differences:

First, only the matrix verb can take an aspect marker or be negated. In other words, the domains of circumstances and purpose dependents differ and offer a more restricted list of fields than the verbal domain. If the relation between *kaihui* and *taolun* is *circumstance* then we have the word order in (13). If however, the relation is *purpose*, we have (14)**Erreur ! Source du renvoi introuvable.**:

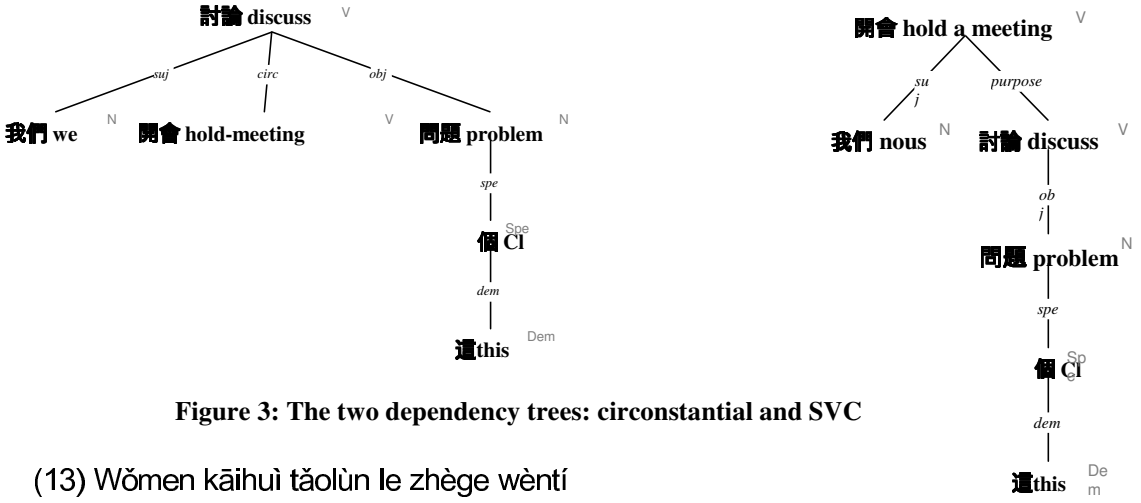


Figure 3: The two dependency trees: circonstantial and SVC

(13) Wǒmen kāihuì tǎolùn le zhège wèntí
'We have discussed this problem during the meeting'
*Wǒmen kāihuì le tǎolùn zhège wèntí

(14) Wǒmen kāihuì le tǎolùn zhège wèntí
'We have held a meeting to discuss this problem'
*Wǒmen kāihuì tǎolùn le zhège wèntí

The constraint on the negation is very similar and straightforward.

Second, we have to consider the possible topicalizations that may affect the word order of such structures. In both cases, the NP "zhège wèntí" '*this problem'* can be topicalized, but (quite obviously) the matrix verb cannot, neither can the purpose verb (without adding lexical material, like prepositions or verb duplication, and completely changing the syntactic structure). Finally, "kāihuì " *'hold a meeting'* can only be topicalized if it is circumstantial[6]. Our grammar generates all and only these word order possibilities. We only need to introduce the following rules for the SVC reading:

[6] Some informants don't accept the topicalization of a bare verb, or find it unnatural. If we add the postposition 時shí to the verb, however, the verbal topicalization becomes generally acceptable. This particle appears at the syntactic level and can be dealt with a small amendment to our grammar adding a constraint on the topic field. We don't want to stress this point here for clarity reasons.

Reduced verbal domain: rvd = verb! Object? and Domain creation rule :(SVC,V,rvd,verb)

These additions to our grammar give a different interesting result: Starting with two different dependency structures, we obtains various word orders, some of them common to the two different dependency structures, attesting that the surface form is ambiguous. We also noticed that all the word orders (but none of the topological trees) generated by the SVC dependency tree can be generated from the circumstantial dependency tree, while the contrary does not hold. This observation seems to suit the preferences of our native speaker informants.

Below we show all possible word orders for the first dependency tree (with the circumstantial dependency, 8 word orders):

8. i[*macro-d* topic[*nd* noun我們we] topic[*nd* dem這this cl個 noun問題problem] micro[*micro-d* verbal[*verbal-d* circ開會hold a meeting verb討論discuss]]]

15. i[*macro-d* topic[*nd* noun我們we] topic[*nd* dem這this cl個 noun問題problem] topic開會hold a meeting micro[*micro-d* verbal[*verbal-d* verb討論discuss]]]

Number of structures with the same word order: 2

---

1. i[*macro-d* micro[*micro-d* subject[*nd* noun我們we] verbal[*verbal-d* circ開會hold a meeting verb討論discuss] object[*nd* dem這this cl個 noun問題problem]]]

6. i[*macro-d* topic[*nd* noun我們we] micro[*micro-d* verbal[*verbal-d* circ開會hold a meeting verb討論discuss] object[*nd* dem這this cl個 noun問題problem]]]

13. i[*macro-d* topic[*nd* noun我們we] topic開會hold a meeting micro[*micro-d* verbal[*verbal-d* verb討論discuss] object[*nd* dem這this cl個 noun問題problem]]]

Number of structures with the same word order: 3

---

16. i[*macro-d* topic[*nd* noun我們we] topic開會hold a meeting topic[*nd* dem這this cl個 noun問題problem] micro[*micro-d* verbal[*verbal-d* verb討論discuss]]]

Number of structures with the same word order: 1

---

2. i[*macro-d* topic[*nd* dem這this cl個 noun問題problem] micro[*micro-d* subject[*nd* noun我們we] verbal[*verbal-d* circ開會hold a meeting verb討論discuss]]]

7. i[*macro-d* topic[*nd* dem這this cl個 noun問題problem] topic[*nd* noun我們we] micro[*micro-d* verbal[*verbal-d* circ開會hold a meeting verb討論discuss]]]

14. i[*macro-d* topic[*nd* dem這this cl個 noun問題problem] topic[*nd* noun我們we] topic開會hold a meeting micro[*micro-d* verbal[*verbal-d* verb討論discuss]]]

Number of structures with the same word order: 3

---

4. i[*macro-d* topic[*nd* dem這this cl個 noun問題problem] topic開會hold a meeting micro[*micro-d* subject[*nd* noun我們we] verbal[*verbal-d* verb討論discuss]]]

10. i[*macro-d* topic[*nd* dem這this cl個 noun問題problem] topic開會hold a meeting topic[*nd* noun我們we] micro[*micro-d* verbal[*verbal-d* verb討論discuss]]]

Number of structures with the same word order: 2

---

3. i[*macro-d* topic開會hold a meeting micro[*micro-d* subject[*nd* noun我們we] verbal[*verbal-d* verb討論discuss] object[*nd* dem這this cl個 noun問題problem]]]

9. i[*macro-d* topic開會hold a meeting topic[*nd* noun我們we] micro[*micro-d* verbal[*verbal-d* verb討論discuss] object[*nd* dem這this cl個 noun問題problem]]]

Number of structures with the same word order: 2

---

12. i[*macro-d* topic開會hold a meeting topic[*nd* noun我們we] topic[*nd* dem這this cl個 noun問題problem] micro[*micro-d* verbal[*verbal-d* verb討論discuss]]]

Number of structures with the same word order: 1

---

5. i[*macro-d* topic開會hold a meeting topic[*nd* dem這this cl個 noun問題problem] micro[*micro-d* subject[*nd* noun我們we] verbal[*verbal-d* verb討論discuss]]]

11. i[*macro-d* topic開會hold a meeting topic[*nd* dem這this cl個 noun問題problem] topic[*nd* noun我們we] micro[*micro-d* verbal[*verbal-d* verb討論discuss]]]

Number of structures with the same word order: 2

This is a list of structures obtained for the second dependency tree (with the SVC, 3 word orders):

5. i[*macro-d* topic[*nd* noun我們we] topic[*nd* dem這this cl個 noun問題problem] micro[*micro-d* verbal[*verbal-d* verb開會hold a meeting] svc[*rvd* verb討論discuss]]]

Number of structures with the same word order: 1

---

1. i[*macro-d* micro[*micro-d* subject[*nd* noun我們we] verbal[*verbal-d* verb開會hold a meeting] svc[*rvd* verb討論discuss object[*nd* dem這this cl個 noun問題problem]]]]

3. i[*macro-d* topic[*nd* noun我們we] micro[*micro-d* verbal[*verbal-d* verb開會hold a meeting] svc[*rvd* verb討論discuss object[*nd* dem這this cl個 noun問題problem]]]]

Number of structures with the same word order: 2

---

2. i[*macro-d* topic[*nd* dem這this cl個 noun問題problem] micro[*micro-d* subject[*nd* noun我們we] verbal[*verbal-d* verb開會hold a meeting] svc[*rvd* verb討論discuss]]]

4. i[*macro-d* topic[*nd* dem這this cl個 noun問題problem] topic[*nd* noun我們we] micro[*micro-d* verbal[*verbal-d* verb開會hold a meeting] svc[*rvd* verb討論discuss]]]

Number of structures with the same word order: 2

6 Conclusion

We have shown that various simple and more complex syntactic phenomena of Chinese find a straightforward formalization in terms of dependency and topology, and thus in the framework of MTT. In spite of some doubts on the usefulness of the commonly used syntactic functions, it is possible to translate into

this type of topological formalization some analyses of syntactic phenomena stemming from different theoretical frameworks, even from "distant" approaches like generativist theories. Contrary to analysis based reasoning that focuses on ambiguities, we believe that this "synthetic" approach explains naturally the underlying linguistic processes. Our approach differs thus in providing the complete set of paraphrases for a given dependency tree, a computation that, as soon as we go beyond the simple examples given in this paper, requires the implementation of the grammar in a computer system.

Our grammar includes some more complex phenomena like for example relative phrases, not presented here for lack of space, and we are working on covering further syntactic details. It would be interesting to explore the connection of this grammar with an implementation of a semantic-syntax interface that could provide the input for our system. On the other end of the pipeline, it remains to be shown that the resulting topological structures have a *raison d'être* in providing a smooth basis for the computation of prosodic groups even for tone language like Chinese.

References

Blanche-Benveniste C. 1990, *Le Français Parlé: Etudes Grammaticales*, CNRS, Paris.

Chao, Yuen Ren. 1968. *A grammar of spoken Chinese.* Berkeley & Los Angeles: University of California press.

Clément, L., K. Gerdes and S. Kahane. 2002. "An LFG-type grammar for German based on topological model", in Miriam Butt and Tracy Holloway King, editors, *Proceedings of the LFG02 Conference*. CSLI Publications, 2002, Stanford

Drach, Erich. 1937. *Grundgedanken der deutschen Satzlehre*, Diesterweg, Frankfurt/M.

Fang, Ji and King, T.H. 2007. "An LFG Chinese Grammar for Machine Use". In. T.H. King and E. M. Bender, eds., Proceedings of the GEAF07 Workshop.

Gerdes K., S. Kahane. 2001. "Word Order in German: A Formal Dependency Grammar Using a Topological Hierarchy" in: *Proceedings ACL 2001*, Toulouse

Gerdes K., S. Kahane. 2007. "Phrasing it differently", in L. Wanner (ed.), *Selected lexical and grammatical issues in the Meaning-Text Theory*, Benja-mins, 297-335.

Her, One-Soon 2003. Chinese inverted constructions within a simplified LMT. *Journal of Chinese Linguistics*, monograph series 19 Lexical-Functional Grammar Analysis of Chinese

Huang, Chu-Ren. 1997. Corpus on web: Introducing the first tagged and balanced Chinese corpus. In: Proceedings of the Annual Conference of the Pacific Neighborhood Consortium.

Li, Charles N, & Thompson, Sandra A.1981. Mandarin Chinese: A functional Reference Grammar. University of California press.

Liu, Haitao. 2007. "Dependency Relations and Dependency Distance: a statistical view based on Treebank". *Proceedings or the Third International Conference on Meaning Text Theory (MTT)*, Klagenfurt, Austria.

LaPolla, Randy J. 1993. "Arguments against 'subject' and 'direct object' as viable concepts in Chinese", in *Bulletin of the Institute of History and Philology* 63.4:759-813.

Mel'čuk, Igor A. 1988. *Dependency Syntax: Theory and Practice*. SUNY Press, Albany, NY.

Mel'čuk, Igor A. 2001. Communicative Organization in Natural Language: The Semantic-Communicative Structure of Sentences. John Benjamins, Amsterdam.

Paul, Waltraud 2005. The "serial verb construction" in Chinese: A Gordian knot. In. Oyharçabal B. & Paul W. 2005. Proceedings of the workshop La notion de « construction verbale en série » est-elle opératoire ? December 9, 2004, Ehess, Paris.

Wu, Ching-huei Teresa 2002. Serial Verb Construction and Verbal Compounding. In Sze-Wing Tang & Chen-Sheng Luther Liu, eds., On the Formal Way to Chinese Languages. CSLI, Stanford, California.

Les dependants syntaxiques de l'adjectif en français : vers un inventaire des relations syntaxiques de surface

Sebastien Marengo
Observatoire de linguistique Sens-Texte
Département de linguistique et de traduction
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal (Québec) H3C 3J7
Canada
sebastien.marengo@umontreal.ca

Resume

Cet exposé présente les premiers résultats d'une recherche en cours sur les dépendants syntaxiques de l'adjectif en français. Dans l'optique de la Théorie Sens-Texte, il défend l'idée selon laquelle il y a tout intérêt à préciser dans le dictionnaire les propriétés syntaxiques des dépendants de l'adjectif, par le principe des relations syntaxiques de surface. Je présente d'abord six propriétés des dépendants dont il faudra rendre compte à l'aide des relations syntaxiques de surface : actant syntaxique profond correspondant, cliticisation, relativation, clivage, dislocation à gauche, position linéaire. Puis, à partir d'un échantillon d'adjectifs, je montre que les valeurs attendues pour certaines de ces propriétés sont moins courantes qu'on ne le pense. Le même échantillon permet enfin d'observer des combinaisons de valeurs pour ces propriétés, autrement dit, d'esquisser quelques relations syntaxiques de surface.

1 Introduction

Les dépendants syntaxiques de l'adjectif ont été moins étudiés que ceux du verbe. Les travaux où sont abordés les dépendants de l'adjectif en français, ou leurs équivalents dans d'autres langues, se concentrent sur certains aspects du domaine :

- Les dépendants syntaxiques les plus étudiés sont de nature infinitivale ou propositionnelle. On peut citer les dépendants syntaxiques des adjectifs modaux (*Luc est facile* ***à convaincre***), ceux des adjectifs « orientés-agent » ou « de qualités morales » (*Paul est gentil* ***de nous aider***), ou encore ceux des adjectifs « psychologiques » (*Léa est contente* ***d'avoir gagné***, *Marie est heureuse* ***que tu aies réussi***). Ces types sont analysés notamment dans les travaux de Gaatone (1972), Bouillon (1996), Riegel (1997), Meunier (1999), Bennis (2000), Léard & Bürgi (2000), Marengo (2002), Léard & Marengo (2005, à paraître), Léger (2006). Outre le type *gentil* ***de la part de Paul***, les dépendants syntaxiques nominaux ont moins retenu l'attention : *rapide* ***des pieds*** (Salles, 1998), *bon* ***au tennis*** (cf. Tucker, 1998), *large* ***de trois mètres*** (cf. Schwarzschild, 2005). L'étude la plus générale reste celle de Picabia (1978), mais la priorité a été donnée « à la description des compléments complétifs » (p. 108).
- L'accent est mis sur certaines des propriétés syntaxiques associées aux dépendants. La propriété la plus couramment vérifiée est la cliticisation (*Luc est facile* ***à convaincre*** *~ *Luc* ***y*** *est facile*). On trouve parfois des observations sur le clivage (*Paul est content* ***d'avoir gagné*** *~ C'est* ***d'avoir gagné*** *que Paul est content*) et sur différents types d'extraction (*Jean a été gentil* ***d'offrir un bijou à Marie*** *~ **<u>***Qu'est-ce que***</u> *Jean a été gentil* ***d'offrir à Marie*** *?*). Il est très rare que les auteurs véri-

fient la dislocation à gauche (***Pour Paul**, cette entreprise est risquée*) ou la relativation (*la fille **dont** Paul est amoureux*), cette dernière étant limitée aux dépendants syntaxiques de type nominal.

Un autre enjeu important concerne la manière de rendre compte des données. Les travaux existants sont difficilement compatibles avec l'approche Sens-Texte :

- Lorsque les auteurs parlent d'un *adjectif*, ils font référence à un vocable et non à une lexie. Il convient donc d'être prudent quant aux observations émises sur les dépendants syntaxiques, car les « constructions » ou les « emplois » distingués ne recoupent pas nécessairement l'inventaire des lexies. Par exemple, dans *Paul est sûr **de se faire mal***, le dépendant syntaxique de l'adjectif autorise apparemment la cliticisation : *Paul **en** est sûr*. Mais en fait, la phrase est ambiguë et peut mettre en jeu une autre lexie du vocable SÛR (cf. Mel'čuk et al., 1999 : 322) : ‹Que Paul va se faire mal est sûr›. Cette fois-ci, la cliticisation n'est pas possible : **Paul **en** est sûr*.
- Il manque parfois des indications claires sur le statut, au sein de représentations sémantiques ou syntaxiques formelles, des expressions étudiées. Certains auteurs cherchent à déterminer s'ils ont affaire à des *arguments* ou *compléments* de l'adjectif, sans qu'on sache si ces termes font référence à la sémantique, à la syntaxe ou aux deux en même temps. Qui plus est, une grande importance est accordée aux tests syntaxiques. On s'attend par exemple à ce qu'un argument second ou complément de l'adjectif autorise le clivage et la cliticisation (1). Si ces tests sont négatifs, une expression a peu de chances d'être considérée comme un argument second ou complément de l'adjectif, et son statut devra être précisé (2). Vraisemblablement, on exclurait ainsi des actants sémantiques ou des dépendants syntaxiques de certains adjectifs, sous prétexte qu'on n'observe pas les comportements attendus. C'était peut-être le cas en (2), et d'autres exemples vont dans le même sens (3).

(1) *Paul est amoureux **de Marie**. ~ C'est **de Marie** que Paul est amoureux. ~ Paul **en** est amoureux.*

(2) *Paul est gentil **de nous aider**. ~ *C'est **de nous aider** que Paul est gentil. ~ *Paul **en** est gentil.*

(3) *La rue est large **de trois mètres**. ~ *C'est **de trois mètres** que la rue est large. ~ *La rue **en** est large.*

Pour toutes ces raisons, il m'a paru souhaitable d'étudier les dépendants syntaxiques de l'adjectif dans la perspective de la Théorie Sens-Texte. Celle-ci permet d'indiquer à même le dictionnaire les propriétés associées aux dépendants syntaxiques, grâce au concept de relation syntaxique de surface (RelSyntS). Les principes-guides pour définir les RelSyntS d'une langue ont été exposés par Iordanskaja & Mel'čuk (2009). Les auteurs appliquent leur méthode au français, en dégageant un inventaire de RelSyntS pour les dépendants syntaxiques contrôlés par la valence du verbe. Mon objectif à long terme est de parvenir à un inventaire comparable pour les dépendants syntaxiques de l'adjectif. Ceux-ci montrent en effet des différences notables par rapport aux dépendants du verbe.

Je vais présenter ici le point de départ de mes recherches. Après les considérations d'usage sur les adjectifs et les RelSyntS (section 2), j'exploiterai un échantillon d'adjectifs pour dégager quelques observations sur les propriétés des dépendants syntaxiques (section 3). Le même échantillon permettra d'esquisser un premier ensemble de RelSyntS (section 4). Je conclurai sur des aspects à développer pour la suite du travail (section 5).

2 L'adjectif, ses dependants syntaxiques et les relations syntaxiques de surface

Les grammaires du français classent en général les dépendants syntaxiques de l'adjectif selon leur catégorie canonique : adverbe (***très** large*) ; nom introduit par une préposition ou parfois par QUE (*amoureux **de sa voisine***, *large **en diable***, *meilleur **que Luc***) ; infinitif introduit par une préposition (*fier **d'avoir gagné***, *fou **à lier***) ; proposition introduite par la conjonction QUE ou l'une de ses variantes (*contente **qu'il fasse***

beau, *attentif* ***à ce que tout se passe bien***, *ravi* ***de ce que tu sois là***). Il est possible aussi que certains adjectifs acceptent un autre adjectif comme dépendant syntaxique : *amoureux* ***fou***, *réputé* ***acceptable***[1].

Une telle classification superficielle n'indique pas quels dépendants syntaxiques sont contrôlés par la valence de l'adjectif — autrement dit, lesquels correspondent à des actants syntaxiques profonds (ASyntP). Suivant la définition des types d'actants retenue dans le cadre de la Théorie Sens-Texte (Mel'čuk, 2004a, 2004b), les adverbes seront toujours considérés comme des modificateurs ; les noms et infinitifs pourront correspondre à des ASyntP ou être des modificateurs ; les propositions correspondront toujours à des ASyntP, la distinction avec les « circonstancielles » étant assurée par la conjonction. Pour la suite de l'exposé, je vais me concentrer sur les dépendants syntaxiques contrôlés par la valence de l'adjectif. Il ne sera donc pas question des modificateurs, sauf ponctuellement, quand ils permettent d'observer la position linéaire des dépendants syntaxiques valenciels (*large d'épaules en diable* ~ ?*large en diable d'épaules*).

Par définition, les adjectifs n'ont pas de I^{er} ASyntP (cf. Mel'čuk, 2004a : 54). L'actant sémantique (ASém) qu'on attendrait dans ce rôle correspondra plutôt au gouverneur syntaxique de l'adjectif (fonction épithète : *la voiture rouge*) ou à l'un des ASyntP du verbe qui gouverne l'adjectif (fonction attribut : *La voiture est rouge*) ; cet élément pourrait être désigné comme le *support* de l'adjectif. La numérotation des ASyntP commencera donc par le chiffre II. Alors que les relations syntaxiques profondes (RelSyntP) se veulent universelles, les RelSyntS sont spécifiques à chaque langue. Une RelSyntS **r** entre un gouverneur (l'adjectif) et son dépendant sera donc le lieu de préciser, pour le français, l'ensemble des propriétés syntaxiques associées au dépendant : possibilité de cliticisation, de clivage, etc.

Selon Iordanskaja & Mel'čuk (2009), une RelSyntS doit répondre à deux types d'exigences : une exigence d'ordre linguistique et une série d'exigences formelles.

Sur le plan linguistique, les dépendants d'une RelSyntS doivent posséder des propriétés similaires en ce qui a trait à la structure syntaxique profonde, à la structure syntaxique de surface et à la structure morphologique profonde. Le nom **r** d'une RelSyntS précisera une famille de constructions syntaxiques de surface qui possèdent des propriétés linguistiques suffisamment semblables, autrement dit, qui présentent des « ressemblances de famille ».

Chaque RelSyntS sera ainsi caractérisée par des propriétés spécifiques du dépendant. En ce qui nous concerne, il y aura moins de propriétés pertinentes que si le gouverneur était un verbe. On peut en recenser au moins six. La première a trait aux rapports avec le niveau syntaxique profond et, corollairement, avec le niveau sémantique (propriété syntactico-sémantique) :

1. Le fait de correspondre à un ASyntP particulier de l'adjectif : *Paul est amoureux* ***de sa voisine*** [= II^{e} ASyntP] ; *Jean est redevable* ***à Paul*** [= II^{e} ASyntP] ***de trois dollars*** [= III^{e} ASyntP].

Les trois propriétés suivantes concernent directement la structure syntaxique de surface (propriétés purement syntaxiques) :

2. Cliticisation : *Paul* ***en*** *est amoureux*.
3. Relativation (applicable pour les noms) : *la fille* ***dont*** *Paul est amoureux*.
4. Clivage (applicable pour les noms et les infinitifs) : *C'est* ***de Marie*** *que Paul est amoureux*.

Les deux dernières propriétés visent l'expression des dépendants dans la structure morphologique profonde, sous l'angle de la linéarisation et de la prosodisation (propriétés syntactico-morphologiques) :

5. Dislocation à gauche (applicable pour les noms et les infinitifs) : ?***De Marie****, Paul est amoureux depuis longtemps*.
6. Position linéaire (non applicable pour les clitiques, les pronoms relatifs, les éléments clivés, les éléments disloqués et les éléments antéposés pour interrogation ou subordination). La position des

[1] On trouve aussi des noms et des infinitifs sans préposition : *ouvert* ***la nuit***, *réputé* ***avoir du talent***. Les infinitifs sans préposition apparaissent notamment avec des lexies qui sont moins clairement des adjectifs, étant donné leur compatibilité avec l'impersonnel : *Il est censé* ***pleuvoir***, *Il est présumé* ***exister des problèmes dans cette entreprise***, *Il est supposé* ***faire beau demain***.

dépendants syntaxiques valenciels par rapport à l'adjectif est fixe : ils sont tous postposés[2]. Ce sera donc l'insertion d'un codépendant entre l'adjectif et le dépendant étudié qui sera déterminante : *large **d'épaules** en diable* ~ *?large en diable **d'épaules***.

Comme on l'aura constaté, quatre propriétés supposent que l'adjectif est gouverné par un verbe : cliticisation, relativation, clivage, dislocation à gauche. Si l'adjectif refuse la fonction attribut, ces propriétés seront sans objet.

Quant aux exigences formelles identifiées par Iordanskaja & Mel'čuk, il faut notamment que toute RelSyntS possède un dépendant prototypique, c'est-à-dire un dépendant d'une telle classe syntaxique qu'il puisse être utilisé avec tout gouverneur possible pour cette RelSyntS. Par exemple, la RelSyntS « sujet » en français admet toujours un nom (ou un pronom) comme dépendant ; il n'existe pas de verbe pour lequel le sujet ne puisse pas être un nom.

Les RelSyntS contrôlées par la valence d'un adjectif devraient être précisées dans le ou les tableaux de régime de ce dernier, pour chacune des *réalisations*. Je fais référence ici aux différents moyens d'expression d'un ASyntP donné, regroupés dans une même colonne. Par exemple, le II[e] ASyntP de la lexie SÛR**I.1a** possède trois réalisations :

Y = II
1. *de* N 2. *de* V_{inf} 3. *que* PROP obligatoire

Figure 1. Tableau de régime de SÛR**I.1a** (Mel'čuk et al., 1999 : 320)

On trouvera ainsi des exemples comme *sûre **de son succès** ~ sûre **de réussir** ~ sûre **qu'elle réussira***. On pourrait parler de *coréalisations* : une réalisation R_1 et une réalisation R_2 sont des coréalisations si et seulement si elles correspondent au même ASyntP du même régime de la même lexie. Pour continuer avec l'exemple de SÛR**I.1a**, on dirait que les réalisations *de N*, *de V_{inf}* et *que PROP* du II[e] ASyntP sont des coréalisations — ou encore, que chacune de ces réalisations possède deux coréalisations.

Notons qu'une RelSyntS doit effectivement être précisée pour chaque réalisation et non pour l'ASyntP dans son ensemble, puisque les différentes réalisations d'un ASyntP donné peuvent avoir des propriétés bien distinctes. Ces coréalisations ne seront donc pas nécessairement couvertes par la même RelSyntS, bien que la chose soit courante.

3 Les proprietes syntaxiques des dependants : quelques observations

Lorsqu'on souhaite étudier à grande échelle les dépendants syntaxiques de l'adjectif, on doit traiter un très grand nombre d'informations, et il est utile de se doter d'un outil pour stocker celles-ci. J'ai donc créé une base de données qui permet d'indiquer les propriétés syntaxiques associées aux réalisations. Elle reprend la structure générale du DiCo (cf. Mel'čuk et al, 1995 : 211–223 ; Jousse & Polguère, 2005), qui indique notamment l'ASyntP correspondant. S'y ajoutent des champs relatifs aux propriétés purement syntaxiques (cliticisation, relativation, clivage) et syntactico-morphologiques (dislocation à gauche, position

[2] Notons toutefois qu'un adjectif et son dépendant en QUE peuvent être disposés de part et d'autre du nom. C'est le cas pour AUTRE, MÊME, MEILLEUR, PIRE et MOINDRE : *Paul a une autre voiture **que Luc***. Par ailleurs, lorsqu'une lexie admet aussi bien l'antéposition que la postposition par rapport à son gouverneur, les principes-guides imposent de postuler deux RelSyntS en conséquence (Iordanskaja & Mel'čuk, 2009 : 152). Cela vaut pour plusieurs « adverbes » (***physiquement** résistant ~ résistant **physiquement***) et adjectifs (*une **énorme** maison ~ une maison **énorme***). De telles lexies n'étant pas sélectionnées par la valence de leur gouverneur syntaxique, il serait crucial d'ajouter, dans leurs articles de dictionnaire, des informations sur leur valence passive. De manière analogue, certains adjectifs admettent un infinitif et/ou une proposition comme sujet (*Chanter est agréable*, *Qu'il parte est surprenant*) ; bien qu'un tel phénomène ne concerne pas le régime de l'adjectif à proprement parler, il devrait être consigné dans son article de dictionnaire.

linéaire). Ces dernières sont dégagées à partir de plusieurs exemples, eux aussi conservés dans la base de données.

Au moment d'écrire ces lignes, la base de données contient un ensemble d'adjectifs dont les dépendants syntaxiques sont bien caractérisés : il s'agit des adjectifs régissants qui figurent dans le *Dictionnaire explicatif et combinatoire* (DEC, Mel'čuk et al., 1984–1999) et dans le DiCouèbe. Cela correspond à 23 vocables, 40 lexies régissantes et 77 réalisations[3], le tout illustré par 785 exemples. Plusieurs de ces exemples sont ceux du DEC et du DiCouèbe eux-mêmes, mais la plupart sont extraits de la base Frantext[4] ; quelques-uns proviennent du Web ; d'autres enfin ont été créés de toutes pièces, parfois en modifiant un exemple donné par le DEC ou le DiCouèbe.

Malgré son caractère limité, l'échantillon permet quelques observations intéressantes en ce qui a trait aux propriétés des dépendants syntaxiques. Afin de respecter l'espace imparti, je me concentrerai sur l'ASyntP correspondant, la cliticisation, la relativation et le clivage.

3.1 Actant syntaxique profond correspondant

Comme on peut s'y attendre, la plupart des 77 réalisations de l'échantillon correspondent au II^{e} ASyntP de la lexie : c'est le cas pour 67 d'entre elles. Quelques-unes correspondent au III^{e} ASyntP : elles sont au nombre de neuf. On trouve même une lexie dotée d'un IV^{e} ASyntP : *la dette payable par ce pays à l'Angleterre* ***en dollars américains***. On peut s'interroger sur l'existence de lexies pourvues d'un V^{e} ASyntP : *Y louable par X à Z pour la somme W* ***pendant la période T***.

3.2 Cliticisation

L'échantillon montre bien que la cliticisation n'est pas toujours possible pour les dépendants syntaxiques contrôlés par la valence de l'adjectif. Si 24 réalisations la permettent (4a), 14 ne l'autorisent pas (4b). Elle paraît douteuse pour cinq réalisations (4c). Pour les 34 réalisations restantes, elle est considérée comme sans objet : soit une préposition est en jeu et ce n'est ni À ni DE (4d), soit la lexie refuse la fonction attribut.

(4) a. *Je suis fier* ***d'avoir réussi***. ~ *J'****en*** *suis fier.*

b. *Pierre est malade* ***des reins***. ~ **Il* ***en*** *est malade*.

c. *Le gardien était armé* ***de son couteau***. ~ *?Il* ***en*** *était armé*.

d. *Le rôle est casse-gueule* ***pour cette comédienne***. ~ **Le rôle* ***lui*** ⟨***y***, ***en***⟩ *est casse-gueule*.

Il faut cependant noter, même si l'échantillon ne le montre pas, que la cliticisation est quelquefois possible alors que la préposition ne la laisse pas attendre :

(5) a. *Je suis reconnaissant* ***envers Pierre***. *Je* ***lui*** *suis même très reconnaissant*.

b. *Je vous suis reconnaissant* ***pour votre aide***. *Je vous* ***en*** *suis même très reconnaissant*.

c. *Ce stage est utile* ***pour les apprentis***. *Il* ***leur*** *est même très utile*.

[3] En fait, neuf de ces réalisations ont été ajoutées par mes soins, dans la mesure où j'estime qu'elles auraient dû figurer dans les articles de dictionnaire.

[4] Les exemples de Frantext ont été extraits à l'aide d'une « grammaire » que j'ai rédigée. Celle-ci permet de spécifier les éléments recherchés (adjectif, préposition, verbe copule, clitique, pronom relatif…) en tenant compte de leurs éventuelles variations morphologiques, de préciser leur position relative et d'allouer un nombre déterminé d'éléments quelconques pouvant les séparer les uns des autres. Cette démarche a pour objectif d'accélérer la recherche d'exemples pertinents mais ne vise en aucune façon l'exhaustivité ni l'établissement de statistiques.

Cela tient parfois à ce qu'une coréalisation contient la préposition attendue :

(6) a. *Je suis reconnaissant **à Pierre**. Je **lui** suis même très reconnaissant.*

b. *Je vous suis reconnaissant **de votre aide**. Je vous **en** suis même très reconnaissant.*

c. *Ce stage est utile **aux apprentis**. Il **leur** est même très utile.*

Mais le fait que la préposition attendue soit possible n'est pas une condition suffisante :

(7) a. *Je suis étonné **devant son succès**. *J'**en** suis même très étonné.*

b. *Je suis étonné **de son succès**. J'**en** suis même très étonné.*

Ce n'est pas non plus une condition nécessaire (8a, 8b). Picabia (1978 : 73–76) note toutefois que le dédoublement ramène la préposition attendue (8c, 8d).

(8) a. *Il est capital **pour Marie** d'obtenir ce livre. ~ Il **lui** est capital d'obtenir ce livre.*

b. **Il est capital **à Marie** d'obtenir ce livre.*

c. **Il **lui** est capital, **pour Marie**, d'obtenir ce livre.*

d. *Il **lui** est capital, **à Marie**, d'obtenir ce livre.*

Un cas particulier pourrait concerner des ASyntP qui n'apparaîtraient que sous forme de pronom, disjoint ou clitique : *Il est impossible **pour lui** de rester ~ Il **lui** est impossible de rester ~ ?Il est impossible **pour Luc** de rester ~ ?Il est impossible **à Luc** de rester* (noter toutefois *Il **lui** est impossible, **à Luc**, de rester*). Certains exigeraient peut-être même le clitique : *Il **lui** est loisible de rester ~ ?Il est loisible **à Luc** de rester ~ *Il est loisible **à lui** de rester* (noter toutefois *Il **lui** est loisible, **à Luc**, de rester*).

3.3 Relativation

En ce qui concerne la relativation, 33 réalisations sur 77 l'autorisent (9a), 11 l'interdisent (9b) et 12 donnent un résultat douteux (9c). Elle est sans objet pour les 21 réalisations restantes : soit la locution prépositionnelle ne la permet pas (9d), soit on a affaire à un infinitif ou une proposition, soit la lexie refuse la fonction attribut.

(9) a. *Une révolution **dont** il fut encore plus étonné que bien d'autres*

b. **Le rein **dont** il est malade*

c. *?Le pays **par lequel** cette dette est payable*

d. *Léo est irréprochable **en tant que mari**. ~ *Un mari, **en tant que lequel** Léo est irréprochable, …*

Pour l'instant, la propriété de relativation correspond aux types standard : la relative doit dépendre syntaxiquement d'un nom. Mais il serait possible de considérer les types où la relative est introduite par CE :

(10) a. *Ce **dont** Marie est sûre, c'est d'avoir réussi.*

b. *Ce **dont** je suis sûr, c'est que la lecture de Pascal me conduisit à cette atroce hypothèse.*

Il s'agit des « relatives périphrastiques » de Riegel et al. (2001). L'« antécédent » peut être un fait et donc apparaître sous forme d'infinitif ou de proposition. Le test serait pertinent au moins pour les infinitifs, dans la mesure où certains ne permettent pas la relativation :

(10) c. *Je suis bien con* ***de me fatiguer***. ~ **Ce* ***dont*** *je suis bien con, c'est de me fatiguer.*

Il est possible cependant que ce type de relativation ne soit autorisé que si la réalisation infinitive possède une coréalisation nominale. Cela expliquerait le blocage en (10c).

3.4 Clivage

De nos 77 réalisations, 33 permettent le clivage (11a), 12 l'interdisent (11b) et 25 laissent planer un doute (11c). La propriété est sans objet pour les sept autres réalisations, qui correspondent à une proposition (11d) ou pour lesquelles la lexie refuse la fonction attribut.

(11) a. *C'est* ***de sa voiture*** *qu'il est fier.*

b. **C'est* ***de me fatiguer*** *que je suis bien con.*

c. *?C'est* ***de la grippe espagnole*** *que Pierre est malade.*

d. *Je suis étonné* ***qu'elle soit venue***. ~ **C'est* ***qu'elle soit venue*** *que je suis étonné.*

On peut se demander si une réalisation infinitive doit nécessairement posséder une coréalisation nominale pour autoriser le clivage. La base de données ne contient pour l'instant aucun contre-exemple.

4 Vers un inventaire des relations syntaxiques de surface

L'échantillon d'adjectifs contenus dans la base de données permet aussi d'esquisser un premier ensemble de RelSyntS. On peut d'abord prévoir combien de familles de RelSyntS seront nécessaires en fonction du dépendant prototypique. Il suffit de repérer les ASyntP dont toutes les réalisations appartiennent à une même classe syntaxique ; cela inclut les réalisations *uniques*, c'est-à-dire dépourvues de coréalisations. Pour chacune de ces réalisations, qui sont au nombre de 48, la RelSyntS postulée devra nécessairement avoir un dépendant prototypique de la classe syntaxique visée. On peut ensuite définir quelques RelSyntS, en observant les combinaisons possibles pour les valeurs des propriétés syntaxiques.

L'échantillon montre qu'on doit en premier lieu prévoir une famille de RelSyntS dont le dépendant prototypique est un nom, ce qui était prévisible. On trouve en effet 43 réalisations nominales qui sont dépourvues de coréalisations infinitivales ou propositionnelles. Je vais présenter ici trois RelSyntS.

La première peut correspondre au II^{e}, III^{e} ou IV^{e} ASyntP ; elle présente des valeurs positives pour la cliticisation, la relativation, le clivage et l'insertion d'un codépendant, la dislocation à gauche étant légèrement douteuse (12). Cette RelSyntS s'applique assurément à deux réalisations sur les 43, et quatre autres réalisations ont des propriétés similaires.

(12) a. *Je suis sûr* ***de son heure d'arrivée*** [= II].

b. *J'****en*** *suis sûr.*

c. *Son heure d'arrivée,* ***dont*** *je suis sûr, …*

d. *C'est* ***de son heure d'arrivée*** *que je suis sûr.*

e. *?****De son heure d'arrivée****, je suis absolument sûr.*

f. *Je suis sûr à 100 %* ***de son heure d'arrivée****.*

Pour la deuxième RelSyntS, l'ASyntP correspondant est le II[e] ou le III[e]. La cliticisation est sans objet (en raison de la préposition en jeu), mais toutes les autres propriétés sont positives, y compris la dislocation à gauche, ce qui est caractéristique (13). Cette RelSyntS est à coup sûr la mieux représentée : elle s'applique à 18 réalisations sur les 43, plus potentiellement trois autres.

(13) a. *La restauration est casse-gueule **pour les débutants*** [= II].

b. *Les débutants, **pour qui** la restauration est casse-gueule, …*

c. *C'est **pour les débutants** que la restauration est casse-gueule.*

d. ***Pour les débutants**, la restauration est casse-gueule.*

e. *La restauration est casse-gueule en diable **pour les débutants**.*

Avec la troisième RelSyntS, on a affaire au II[e] ASyntP. Seuls le clivage et l'insertion sont positifs (14). Une réalisation sur les 43 est clairement en jeu, et peut-être une deuxième.

(14) a. *Vous êtes malade **du poumon*** [= II].

b. **Vous **en** êtes malade.*

c. **Le poumon **dont** vous êtes malade…*

d. *C'est **du poumon** que vous êtes malade.*

e. ?*Et **du poumon**, vous êtes malade depuis longtemps ?*

f. *Vous êtes malade depuis longtemps **du poumon** ?*

Ces trois RelSyntS peuvent être identifiées temporairement par les lettres A, B et C (Tableau 1). Si les deux premières correspondent à des comportements « normaux », la troisième révèle une différence intéressante par rapport à l'inventaire de RelSyntS pour lesquelles le gouverneur est un verbe (cf. Iordanskaja & Mel'čuk, 2009 : 221) : le clivage est positif alors que la cliticisation est négative.

	A	B	C
1. ASyntP correspondant	II/III/IV	II/III	II
2. Cliticisation	EN	Ø	–
3. Relativation	+	+	–
4. Clivage	+	+	+
5. Dislocation à gauche	?	+	?
6. Insertion	+	+	+

Tableau 1. RelSyntS dont le dépendant prototypique est un nom

On pourrait penser que les ASyntP de l'adjectif peuvent toujours apparaître sous forme de nom. Or, sur les 48 réalisations, trois réalisations infinitivales sont uniques. Deux ont des caractéristiques identiques : elles correspondent au II[e] ASyntP et ne permettent que l'insertion, la relativation « standard » étant sans objet (15).

(15) a. *Certains soins sont sûrs **de figurer dans le contrat**.*

b. **Certains soins **en** sont sûrs.*

c. **C'est **de figurer dans le contrat** que certains soins sont sûrs.*

d. ****De figurer dans le contrat**, certains soins sont sûrs.*

e. *Certains soins sont sûrs désormais **de figurer dans le contrat**.*

Les deux réalisations seront donc couvertes par la même RelSyntS (Tableau 2). Cette fois-ci, il n'y a pas de différence notable par rapport au verbe (cf. Iordanskaja & Mel'čuk, 2009 : 223).

	D
1. ASyntP correspondant	II
2. Cliticisation	–
3. Relativation	s.o.
4. Clivage	–
5. Dislocation à gauche	–
6. Insertion	+

Tableau 2. RelSyntS dont le dépendant prototypique est un infinitif

Plus surprenant encore, l'échantillon contient une lexie dont le II^e^ ASyntP n'apparaît que sous forme de proposition : il s'agit de SÛR**I.2b**. La proposition peut être introduite par QUE ou par une pause : *Sûr **qu'il y a des rats chez Igor** ! ~ Sûr, **il y a des rats chez Igor** !*. Certes, il n'est pas sûr (!) qu'on ait ici affaire à un adjectif. On n'observe ni la dépendance syntaxique à l'égard d'un nom (fonction épithète) ni la dépendance syntaxique à l'égard d'un verbe (fonction attribut). On peut aussi se demander si SÛR gouverne bien la proposition en l'absence de conjonction, ou si ce ne serait pas plutôt l'inverse. Toujours est-il que d'autres lexies ont des caractéristiques semblables : *Bizarre **qu'il soit là***, *Dommage **que tu t'en ailles*** (cf. aussi *Impossible **de lui parler***)[5].

La relativation, le clivage et la dislocation à gauche sont par définition sans objet pour les propositions. Avec SÛR**I.2b**, la cliticisation l'est aussi, mais de manière ponctuelle, en raison du refus de la fonction attribut. Quant à l'insertion, elle est difficilement vérifiable, puisque SÛR**I.2b** n'accepte comme autre dépendant qu'un marqueur de négation (*Pas sûr **qu'il y ait des rats chez Igor***). La RelSyntS se résumera donc comme suit :

	E
1. ASyntP correspondant	II
2. Cliticisation	?
3. Relativation	s.o.
4. Clivage	s.o.
5. Dislocation à gauche	s.o.
6. Insertion	?

Tableau 3. RelSyntS dont le dépendant prototypique est une proposition

[5] Selon le DEC, la lexie SÛR**I.2b** est dérivée de SÛR**I.2a**, qui possède deux ASém : *Il est sûr pour moi* [= X] *qu'il y a des rats chez Igor* [= Y]. Comme l'indiquent les auteurs, il y a deux bonnes raisons de considérer SÛR**I.2b** comme une lexie à part. Primo, contrairement à SÛR**I.2a**, elle « n'admet ni l'expression de la personne pour qui Y est sûr (**Sûr pour moi qu'il va neiger*), ni les modificateurs (**Tout à fait sûr qu'il va neiger*) » (Mel'čuk et al., 1999 : 322). Secundo, les adjectifs de la classe sémantique de SÛR**I.2a** ne donnent pas tous lieu à ce type de polysémie : *Il est douteux qu'il soit là ~ *Douteux qu'il soit là*. On n'a donc pas affaire à une simple ellipse du sujet et de la copule : (*C'est*) *sûr qu'il y a des rats chez Igor*. Quant au statut sémantique et syntaxique de SÛR**I.2b**, il est possible qu'il s'agisse d'un marqueur discursif. Dostie & Lanciault (2008) proposent une telle analyse pour SÉRIEUX, lui aussi d'origine adjectivale : *Sérieux, à Toronto ça fume pas dans les bars*.

5 Conclusion

Au cours des prochaines semaines, la base de données s'enrichira et devrait permettre de dresser un inventaire plus représentatif de RelSyntS. On peut penser que la tâche sera moins complexe que pour les dépendants syntaxiques du verbe, étant donné qu'il y a moins de propriétés pertinentes à observer. En fait, le nombre moins élevé de propriétés peut rapidement devenir un handicap : quand certaines sont sans objet, on risque d'hésiter entre différentes RelSyntS[6]. Les propriétés restantes suffiront-elles à déterminer sans équivoque la RelSyntS en jeu ? Devra-t-on se tourner vers les coréalisations pour trancher ?

D'autres phénomènes, plus particuliers, soulèvent de belles questions lexicographiques, à la fois sur le plan théorique et pratique. Par exemple :

- Il n'est pas toujours facile d'identifier les dépendants syntaxiques de l'adjectif. Quelques adjectifs du français semblent confier au nom la « garde syntaxique » d'un de leurs ASém : *le sport préféré* ***de Paul*** 〈le sport que Paul préfère〉 (cf. Mel'čuk, 2004b : 271, d'après Boguslavskij) ; *la ville natale* ***de Luc*** 〈la ville où Luc est né〉. On le voit d'autant mieux que l'ASém en question peut apparaître sous forme de déterminant possessif : ***son*** *sport préféré* ; ***sa*** *ville natale*. Comment indiquer un tel phénomène dans l'article de dictionnaire ?
- Certains intensificateurs paraissent incompatibles avec des dépendants contrôlés par la valence : *Paul est con* ***comme un panier*** ~ *Paul est con* ***d'avoir fait ça*** ~ ?*Paul est con* ***comme un panier d'avoir fait ça*** ; *Je suis heureux* ***comme un pape*** ~ *Je suis heureux* ***de te voir*** ~ ?*Je suis heureux* ***comme un pape de te voir***. Faut-il y voir un indice de polysémie ou une contrainte sur la combinaison des dépendants syntaxiques ?

Bref, un portrait général des dépendants syntaxiques de l'adjectif en français semble s'imposer. Cela, apparemment, n'a jamais été envisagé, ou, du moins, pas dans l'optique de la Théorie Sens-Texte.

Remerciements

J'aimerais d'abord remercier Alain Polguère, qui m'a encouragé à créer la base de données et m'a consacré beaucoup de son temps pour répondre à mes questions. Je suis également *reconnaissant envers* Jean-Marcel Léard pour ses commentaires et suggestions en matière adjectivale, qui ne datent pas d'hier. Ma gratitude va aussi aux relecteurs anonymes, qui m'ont fait part de leurs observations sur la version initiale de l'exposé. Merci enfin au Conseil de recherches en sciences humaines du Canada (CRSH) pour son appui financier.

Bibliographie

Barrier, Nicolas. 2002. Une MétaGrammaire pour les adjectifs du français. In *Actes du colloque TALN 2002*, 351–357. http://www.loria.fr/projets/JEP-TALN/actes/TALN/posters/Poster06.pdf.

Bennis, Hans. 2000. Adjectives and Argument Structure. In Peter Coopmans, Martin Everaert & Jane Grimshaw (eds), *Lexical Specification and Insertion*, 27–67. Amsterdam/Philadelphia: Benjamins.

Bouillon, Pierrette. 1996. Le lexique génératif : une alternative au traitement de la polysémie. Le cas des adjectifs qui dénotent un état mental. In André Clas, Philippe Thoiron & Henri Béjoint (éds), *Lexicomatique et dictionnairiques*, 359–369. Montréal : AUPELF–UREF.

DiCouèbe. Dictionnaire en ligne de combinatoire du français, Observatoire de linguistique Sens-Texte (OLST), Université de Montréal, http://olst.ling.umontreal.ca/dicouebe/.

Dostie, Gaétane & Lisanne Lanciault. 2008. Changement catégoriel et développement sémantique. De *sérieux* adjectival à *sérieux* discursif dans le parler des jeunes locuteurs québécois. Colloque *Modes langagières dans l'histoire*. Montpellier : Université Paul Valéry. 11 au 13 juin.

[6] Cela survient notamment quand la lexie refuse la fonction attribut, puisque la cliticisation, la relativation, le clivage et la dislocation à gauche sont inapplicables. Par bonheur, les adjectifs régissants qui refusent la fonction attribut ne sont pas très nombreux (cf. Marengo, 2007) : *Marie avait une autre robe* 〈*la même robe*〉 ***que Léa***.

Frantext, ATILF/CNRS, Université de Nancy 2, http://www.frantext.fr.

Gaatone, David. 1972. Facile à dire. *Revue de linguistique romane*, 36(1):129–138.

Iordanskaja, Lidija & Igor Mel'čuk. 2009. Establishing an Inventory of Surface-Syntactic Relations: Valence-Controlled Surface-Syntactic Dependents of the Verb in French. In Alain Polguère & Igor A. Mel'čuk (eds), *Dependency in Linguistic Description*, 151–236. Amsterdam: Benjamins.

Jousse, Anne-Laure & Alain Polguère. 2005. *Le DiCo et sa version DiCouèbe. Document descriptif et manuel d'utilisation*. http://olst.ling.umontreal.ca/dicouebe/DiCoDOC.pdf.

Léard, Jean-Marcel. En préparation. *Grammaire sémantique modulaire : lexique, référence, prédication.*

Léard, Jean-Marcel & Anne Bürgi. 2000. *Tu es naïf de croire que c'est facile à analyser* : catégories et modularité. In *Lexique, Syntaxe et Sémantique : mélanges offerts à Gaston Gross*, 231–241. Besançon : Bulag.

Léard, Jean-Marcel & Sébastien Marengo. 2005. Pour une typologie des compléments adjectivaux : arguments, quasi-arguments et non-arguments. In Jacques François (dir.), *L'adjectif en français et à travers les langues. Actes du colloque international de Caen (28–30 juin 2001)*, 387–402. Caen : Presses universitaires de Caen.

Léard, Jean-Marcel & Sébastien Marengo. À paraître. Le syntagme adjectival et les arguments de l'adjectif. In Anne Abeillé, Annie Delaveau & Danièle Godard (dirs), *Grande grammaire du français*. Paris : Bayard.

Léger, Catherine. 2006. *La complémentation de type phrastique des adjectifs en français*. Thèse de doctorat. Montréal : Université du Québec à Montréal.

Marengo, Sébastien. 2002. *L'adjectif : classification sémantique et structures d'arguments*. Mémoire de maîtrise. Sherbrooke : Université de Sherbrooke.

Marengo, Sébastien. 2007. *L'adjectif non-attribut. Syntaxe et sémantique des adjectifs référentiels*. Thèse de doctorat. Strasbourg : Université Marc Bloch.

Mel'čuk, Igor. 1988. *Dependency Syntax: Theory and Practice*. Albany: State University of New York Press.

Mel'čuk, Igor. 2004a. Actants in semantics and syntax I: actants in semantics. *Linguistics*, 42(1):1–66.

Mel'čuk, Igor. 2004b. Actants in semantics and syntax II: actants in syntax. *Linguistics*, 42(2):247–291.

Mel'čuk, Igor et al. 1984, 1988, 1992, 1999. *Dictionnaire explicatif et combinatoire du français contemporain*, vol. 1–4. Montréal : Presses de l'Université de Montréal.

Mel'čuk, Igor, André Clas & Alain Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve : Duculot.

Mel'čuk, Igor & Nikolaj Pertsov. 1987. *Surface Syntax of English. A Formal Model within the Meaning-Text Framework*. Amsterdam/Philadelphia: Benjamins.

Meunier, Annie. 1999. Une construction complexe *N_0hum être Adj de V^0-inf W* caractéristique de certains adjectifs à sujet humain. *Langages*, 133:12–44.

Picabia, Lélia. 1978. *Les constructions adjectivales en français. Systématique transformationnelle*. Genève : Droz.

Riegel, Martin. 1997. *Il est gentil de nous avoir aidés* ou : à propos de compléments de l'adjectif qui n'en sont pas vraiment. In Georges Kleiber & Martin Riegel (éds), *Les formes du sens*, 355–365. Louvain-la-Neuve : Duculot.

Riegel, Martin, Jean-Christophe Pellat & René Rioul. 2001. *Grammaire méthodique du français*, 6^{e} édition. Paris : Presses Universitaires de France.

Salles, Mathilde. 1998. La construction converse *être un peu lent de la tête mais rapide des pieds. La Linguistique*, 34(1):121–136.

Schwarzschild, Roger. 2005. Measure Phrases as Modifiers of Adjectives. *Recherches linguistiques de Vincennes*, 34:207–228.

Tucker, Gordon H. 1998. The syntax of adjectives: the quality group. *The Lexicogrammar of Adjectives. A Systemic Functional Approach to Lexis*, 61–91. London/New York: Cassell.

C'est la definition de quel mot?
Tester la validité des définitions lexicographiques pour un dictionnaire d'apprentissage

Jasmina Milićević
Dalhousie University & OLST – Université de Montréal
jmilicev@dal.ca

Résumé

Cet article décrit un test effectué dans le but de vérifier la validité d'un sous-ensemble de définitions lexicographiques élaborées pour un dictionnaire d'apprentissage du français langue seconde. Le test consistait à soumettre aux participants les définitions d'une vingtaine de lexies françaises dénotant des sentiments et à leur demander d'identifier les lexies correspondantes. Les résultats du test étaient censés mettre en évidence des lacunes des définitions proposées et – à travers une analyse des difficultés éprouvées par les participants – des pistes à suivre afin de les améliorer. Ces résultats nous ont permis, notamment, de faire une description plus fine des différences entre chaque lexie testée et des lexies sémantiquement apparentées avec lesquelles l'usager du dictionnaire risque de confondre celle-ci.

1 L'objectif et la méthodologie du test

Le présent article décrit un test effectué dans le but de vérifier la validité d'un sous-ensemble de définitions lexicographiques élaborées dans le cadre du projet de dictionnaire *Dire autrement* (Milićević & Hamel 2007, Hamel & Milićević 2007), un dictionnaire électronique d'apprentissage du français langue seconde de niveau intermédiaire-avancé. *Dire autrement* est un dictionnaire de type explicatif et combinatoire « didactisé », c'est-à-dire 1) élaboré selon la méthodologie de la *Lexicologie explicative et combinatoire*, ou la *LEC* (Mel'čuk *et al*. 1995, Mel'čuk 2006), et 2) mis sous une forme plus facilement accessible pour les apprenants (sur les définitions pédagogiques à la LEC voir, notamment, Milićević 2008 et Apresjan *et al*. à paraître ; sur l'encodage convivial des collocations, voir Popovic 2003).

Les définitions de la LEC sont des définitions analytiques, ou paraphrastiques, c'est-à-dire faites par la décomposition du sens de la lexie décrite L en termes de sens plus simples que le sens de L. (Le sens 's_1' est plus simple que le sens 's_2' si 's_1' figure dans la décomposition de 's_2' et que l'inverse n'est pas vrai.)

À titre d'exemple, les définitions du verbe SE TAIRE (*Jean se taisait*) et du nom SECRET (*Je n'ai pas de secrets devant toi*) se présentent comme suit.

DEFINI	DEFINISSANT
X se tait	≡ '`individu` X, qui est censé parler, ne parle pas'
secret de X concernant Z(Y)	≡ '`information` qu'a `individu` X concernant `fait` Z lié à `individu` Y telle que X ne doit pas la communiquer aux autres'

Dans le cas d'une lexie prédicative, le défini est la *forme propositionnelle*, une expression contenant la lexie L et ses actants sémantiques (représentées par des variables X, Y, Z, etc.), et le définissant est la décomposition sémantique proprement dite du sens de L. Les composantes sémantiques imprimées en police spéciale dans les définitions ci-dessus (`individu`, `information` et `fait`) sont des caractériseurs taxinomiques, ou étiquettes sémantiques (sur leur utilisations dans le cadre de la LEC, voir Milićević, 1997 et Polguère, 2004).

Pour vérifier la validité de nos définitions, nous avons eu recours au test de reconnaissance, un test dans lequel il s'agit d'identifier la lexie L en lisant la définition de L. Autrement dit, il faut trouver le mot

à partir de la description de son sens. L'hypothèse à la base du test était qu'une définition adéquate d'une lexie L devrait permettre au locuteur d'identifier L. Une définition adéquate de L est telle que chaque composante en est nécessaire et l'ensemble de composantes est suffisant pour couvrir tous les emplois de L ; c'est ce qu'on connaît sous le nom de *principe d'adéquation*.

Les définitions dans un dictionnaire d'apprentissage doivent en outre être conformes au principe de convivialité. Ceci veut dire que leur format – la structure et le langage des définitions – doit être adapté aux besoins d'apprenants (*grosso modo*, leur niveau d'apprentissage). Nos définitions sont légèrement didactisées par rapport aux définitions standard de la LEC, notamment celles qu'on trouve dans le *Dictionnaire explicatif et combinatoire du français contemporain*, ou DECFC (Mel'čuk *et al*. 1984-1988-1992-1999) ; une comparaison des deux formats de définition sera donnée plus loin. Notre test ne tenait pas compte de l'aspect convivialité des définitions (il n'était pas conçu pour cela). Il est clair, cependant, que la « lisibilité » des définitions dans un dictionnaire d'apprentissage doit être prise en compte lorsqu'on juge de leur adéquation.

Insistons que l'identification réussie n'est pas LA preuve de l'adéquation : il ne s'agit que d'une indication parmi d'autres. (Ainsi, le test de reconnaissance ne permet d'évaluer qu'un aspect de la validité des définitions ; voir immédiatement ci-dessous). Pour illustrer ce fait, considérons les deux définitions suivantes.

RANCUNE (*Nouveau Petit Robert*) Souvenir tenace qu'on garde d'une offense, d'un préjudice, avec de l'hostilité et le désir de vengeance.

Cette définition permet d'identifier assez facilement la lexie visée ; pourtant, son adéquation peut être remise en question. Notamment, la composante 'souvenir', qui est une instance de 'phénomène psychologique', n'est pas le genre prochain de RANCUNE, qui, elle, est une instance de 'sentiment'. On peut également se demander si les composantes 'vengeance' et 'hostilité' sont nécessaires. L'acceptabilité de la phrase *Je lui garde la rancune pour ce qu'il m'a fait, mais je ne cherche pas de vengeance* indique que la rancune ne pousse pas nécessairement à la vengeance (et que, donc, la composante correspondante ne doit pas figurer dans la définition de cette lexie). La situation est moins claire avec la composante 'hostilité' (si l'expression ?*rancune sans hostilité* a l'air bizarre, la disjonction *rancune ou hostilité* semble normale), mais nous somme plutôt d'avis que la rancune ne se manifeste pas nécessairement dans le comportement de la personne qui la ressent. Garder la rancune à quelqu'un c'est lui en vouloir de façon permanente pour ce qu'il nous a fait et vouloir qu'il lui arrive quelque chose de semblable (cf. la locution GARDER UNE DENT, qui est la verbalisation exacte de RANCUNE) ; pour notre définition de RANCUNE, voir plus loin, la sous-section 3.1.

AFRAID (Wierzbicka 1992 : 178)
X thinks something like this:
Something bad can happen.
I do not want this.
I want to do something because of this.
I do not know what I can do.
Because of this, X feels something bad.

Cette définition présente le cas inverse : la reconnaissance de la lexie correspondante est moins facile (s'agit-il de la lexie AFRAID, WORRIED ou APPREHENSIVE, par exemple ?), mais son adéquation ne peut pas être remise en question aussi facilement que celle de la définition du *Petit Robert* ci-dessous.

La reconnaissance de la lexie par la lecture de sa définition (à l'aide de l'intuition linguistique, des connaissances sur la langue[1] et de la logique générale) ne prouve donc pas que la définition est bonne : elle démontre seulement que la définition est suffisante pour établir les distinctions nécessaires (entre L et les lexies apparentés à L). Il s'agit, plus précisément, des deux types suivants de distinctions :

[1] Nous pensons ici aux connaissances spécialisées sur la langue, c'est-à-dire les connaissances en linguistique, que nous opposons à la connaissance (= maîtrise) de la langue, cette dernière ne présupposant pas les connaissances en linguistique.

- Distinctions entre la lexie L et ses quasi-synonymes.

Les quasi-synonymes sont des lexies qui partagent des composantes sémantiques importantes et sont substituables dans au moins quelques contextes ; par exemple, RESPECT#1 [≈ 'X croit que Z(Y) a une grande valeur sociale ou morale et que, à cause de cela, X doit prendre en considération les opinions de Y'] et ADMIRATION [≈ 'X respecte Z(Y) et voudrait être comme Y']. Cf. aussi les séries suivantes de quasi-synonymes : MÉPRIS ~ DÉDAIN ~ IRRESPECT ; MÉCONTENTEMENT ~ INSATISFACTION ~ DÉPLAISIR ; HONTE ~ DÉSHONNEUR ~ HUMILIATION.

- Distinctions entre L et ses L proches.

Ce sont des lexies qui, sans être des quasi-synonymes (elles ne sont pas substituables en contexte), sont sémantiquement apparentées. Comme un exemple de lexies de sens proche, on peut citer MÉPRIS (≈ 'X croit que (Z)Y n'a pas de valeur sociale ou morale et se croit meilleur que Y' ; Y est un individu) et DÉGOÛT (≈ 'X perçoit Y comme très déplaisant et veut l'éviter' ; Y peut être n'importe quoi) ; il y a dans le mépris quelque chose de physiquement désagréable pour X (quelque chose qui se rapproche du dégoût), mais cette parenté est à notre avis trop lointaine et ne devrait pas être reflétée dans les définitions des lexies correspondantes. Cf. aussi les séries suivantes : MÉCONTENTEMENT ~ DÉCEPTION ~ FRUSTRATION ; HONTE ~ GÊNE ; PEINE ~ DÉPRIME ~ GÊNE ~ MALAISE.

Nous nous attendions à ce que notre test mette en évidence des lacunes des définitions proposées et – à travers une analyse des difficultés éprouvées par les participants dans la reconnaissance des lexies correspondantes – des pistes à suivre afin de les améliorer. Notamment, la confusion entre L et un de ses quasi-synonymes indiquerait des problèmes (relativement) mineurs avec la définition de L, qu'il s'agirait alors de peaufiner ; la confusion entre L et un de ses sens proches signalerait des problèmes plus graves avec la définition de L et la nécessité d'une correction plus sérieuse de cette dernière.

Nous voulions également tester l'impact des connaissances en linguistique sur la reconnaissance des lexies, l'hypothèse étant que les connaissances en linguistique (plus précisément, les connaissances en sémantique/lexicologie et tout particulièrement la familiarité avec le formalisme d'écriture des définitions) facilitent la reconnaissance.

2 Le contenu du test et les participants

Le test consistait à identifier, en lisant leurs définitions respectives, les lexies suivantes :

- 17 lexies du champ sémantique 'sentiments':[2]
  COLÈRE#1, HAINE#I.a (envers qqn), HAINE#I.b (envers qqch), HONTE, HOSTILITÉ#1, INDIGNATION, IRRITATION, MÉCONTENTEMENT, MÉPRIS#1 (pour qqn), PEINE, PEUR, RANCUNE, REGRET#I.1 (de qqch), REGRET#I.2 (d'avoir fait qqch), REMORDS, REPENTIR, RESPECT#1 (pour qqn).
- 1 lexie « hors champ » :
  HOSTILITÉ#2 **pl. tant** 'combats armés …', liée par polysémie à HOSTILITÉ#1 'sentiment négatif…'.

Ces lexies appartiennent toutes au français courant (même si elles n'ont pas la même fréquence dans les textes) et sont censées être connues de tout locuteur adulte du français.

Les lexies dénotant les sentiments ont jusqu'à trois actants sémantiques :

1) l'individu X qui éprouve le sentiment;

2) la situation Y (possiblement causée par l'individu X lui-même ou par l'individu Z) que X évalue comme bonne/mauvaise, desirable/indésirable, etc., pour X, cette evaluation causant le sentiment de X;

3) l'individu Z qui est la « source » et/ou la « cible » du sentiment de X.

[2] Plus précisément, le champ sémantique en question est constitué des lexies dénotant les sentiments (AMOUR, HAINE, JOIE, TRISTESSE), les émotions (COUP DE FOUDRE, ÉMOI) et les attitudes émotionnelles (RESPECT, ADMIRATION). Il s'agit d'un champs vaste, avec de multiples liens entre lexies et un taux de polysémie élevé – donc, difficile du point de vue de l'acquisition et de la description lexicographique. Voici, à titre d'illustration, trois schémas de polysémie frequents dans le champ en cause : « sentiment ~ objet du sentiment » (JOIE, ADMIRATION), « sentiment ~ manifestation/instance du sentiment » (COLÈRE, PEUR) et « sentiment envers qqn ~ sentiment envers qqch » (AMOUR , HAINE).

Les définitions des lexies dénotant les sentiments sont élaborées selon le même schéma et contiennent ce qu'on appelle des *blocs de définition standard* (voir Iordanskaja & Mel'čuk 1990, Apresjan & Apresjan 1995, ainsi que Wierzbicka 1992 et 1999) :

1) la caractérisation du sentiment (plaisant, déplaisant, fort, etc.) ;

2) l'évaluation, par X, de la situation Y/son participant Z qui cause le sentiment ;

3) optionnelement, la réaction de X au sentiment ;

4) la composante « tel que », qui indique la nature commune du sentiment décrit (cette composante est nécessaire parce que nous ne pouvons pas appréhender les sentiments des autres que par une comparaison avec ce que nous ressentons dans des situations semblables).

Voici, à titre d'illustration, la définition de la lexie PEUR (*Tous les enfants ont peur de l'obscurité, du loup, des monstres, des voleurs*) :

FORME PROPOSITIONNELLE	~ de l'individu X DE Y
CARACTERISATION DU SENTIMENT	'Sentiment négatif de X
EVALUATION DE LA SITUATION	causé par le fait suivant :
(QUI CAUSE LE SENTIMENT)	X perçoit ou imagine Y comme dangereux pour X ;
	ceci cause chez X une forte envie d'éviter Y.
REACTION AU SENTIMENT	Si ce sentiment de X est suffisamment intense,
	il peut causer que X perde le contôle de lui-même.
COMPOSANTE « TEL QUE »	Ce sentiment est le sentiment
	que les gens ont normalement dans des situations semblables.'

La plupart des définitions utilisées pour le test ont été reprises du DECFC. Dans beacoup de cas, la définition reprise a été remaniée (= le contenu en a été modifié) indépendemment de la mise en forme « apprenant ». A titre de comparaison, la définition de PEUR dans le DECFC II: 276 se présente comme suit :

Peur de X de Y ≡ 'Émotion désagréable de X causée par le fait suivant : X croit que l'événement (lié à) Y (concernant l'être Y précieux pour X) qui lui est indésirable est très probable et que X n'est pas capable de s'opposer à Y, et X veut échapper à Y; cette emotion est telle qu'en augmentant, elle cause que X perde la maîtrise de lui-même; elle est celle qu'on a normalement dans de pareilles situations.'[3]

Neuf personnes ont participé au test, dont huit francophones (1 Québécoise, 3 Français, 3 Acadiens, 1 Camerounais) et un russophone ayant une excellente maîtrise du français. Ces personnes comptaient 3 linguistes-professeurs d'université, 1 doctorant en linguistique, 2 doctorants en études françaises-filière lingustique, 1 étudiant du premier cycle en études françaises-filière lingustique, et 2 personnes sans formation en linguistique. Il s'agissait donc de cinq variétés du français, dont une non native, et de niveaux de connaissances en linguistique très différents ; en outre, on avait affaire à des niveaux de connaissance DE la langue différents (cette connaissance étant plus difficile à évaluer/quantifier dans un test comme le nôtre).

3 Les résultats du test

3.1 Les résultats par lexies

Dans cette sous section, nous indiquons les informations suivantes : le sommaire des réponses (Tableau 1), l'ensemble des réponses (Tableau 2), la distribution des réponses par lexie (Tableau 3), les quasi-synonymes et les sens proches des lexies visées trouvés dans les réponses (Tableau 4).

[3] Nous traitons la lexie PEUR comme une instance de sentiment – plutôt que d'émotion – en prenant en compte les donnés sur la fréquence des collocations *sentiment de peur* (44. 400 d'occurrences sur Google) et *émotion de peur* (4. 200 d'occurrences).

Aucune réponse	10
Réponses correctes	71
Réponses incorrectes • **Quasi-synonymes de L** • **Sens proches de L** • "Off mark"	81 **24** **20** 37
Total des réponses	*152*

Tableau 1 : Sommaire des réponses

Comme on peut le constater, 81 réponses, soit plus que la moitié, sont incorrectes. Parmi celles-ci, plus que la moitié sont des quasi-synonymes et des sens proches des lexies visées, distribuées de façon (presque) égale – 24 et 20, respectivement. Les 37 réponses complètement ratées (= « off mark ») sont de deux variétés.

Les réponses « off-mark » sont pour la plupart constitués des lexies sémantiquement plus éloignées de la lexie visée L que ses quasi-synonymes et sens proches, mais qui

1) partagent avec L certaines composantes, possiblement dans une position différente dans la définition (par ex., HOSTILITÉ#1 ['désir de faire mal à Y'] et HAINE#I.a ['désir que qqch de mauvais arrive à Y'] ; HONTE ['sentiment … 'X considère Y(X) comme inadéquat, ce qui le pousse à **cacher** l'existence de Y'] ~ SECRET ['information sur Y que X est censé **cacher** des autres']) ou

2) sont pragmatiquement reliées à L (par ex., on peut rapprocher RESPECT#1 et AMOUR grâce au fait que les deux sentiments, sans s'impliquer mutuellement, « vont ensemble » dans la réalité ; REGRET et NOSTALGIE se rapprochent par le fait d'être orientés vers des expériences plaisantes révolues, quoique de nature différente).

Un petit nombre de réponses « off mark » est constitué de lexies dont le sens n'a aucun lien avec le sens de la lexie visé (par ex., HAINE#I.a et DÉCEPTION).

	1	2	3	4	5	6	7	8	9	
COLÈRE#1	√	**haine**	√	**ressentiment**	vengeance	vengeance	vengeance	√	√	4/9
HAINE#I.a	mépris	?	déception	√	**malveillance**	dégoût	√	√	√	4/9
HAINE#I.b	√	?	√	désapprob-ation	**intolérance**	irritation	désespoir	**aversion**	√	3/9
HONTE	√	√	impudeur	gêne	√	secret	√	√	√	6/9
HOSTILITÉ#1	**animosité**	√	indignation	irritation	**antiphatie**	haine	désaccord	**antiphatie**	√	2/9
HOSTILITÉ#2	√	**guerre**	**guerre**	**guerre**	**différend**	**guerre**	**guerre**	**guerre**	√	2/9
INDIGNATION	dégoût	?	**réprobation**	**révolte**	regret	crime	justice	dégoût	√	1/9
IRRITATION	√	√	√	**enerve-ment**	**colère**	?	angoisse	anxiété	√	4/9
MÉCONTEN-TEMENT	√	**déception**	**déception**	**déception**	souffrance	rancune	**Insatis-faction**	**frustration**	√	2/9
MÉPRIS#1	indignation	**dégoût**	√	√	**dédain**	jalousie	√	désapproba-	√	4/9
PEINE	désespoir	√	**mal**	**déprime**	**déprime**	?	**gêne**	**malaise**	?	1/9
PEUR	√	√	√	√	√	?	√	√	√	8/9
RANCUNE	√	√	√	√	vengeance	√	√	√	√	8/9
REGRET#I.1	peine	**nostalgie**	**nostalgie**	**nostalgie**	frustration	amour	√	√	√	3/9
REGRET#I.2	√	√	√	√	?	revanche	√	√	√	7/9
REMORDS	√	**regret**	√	**culpabilité**	dépréciation	**regret**	√	**regret**	√	4/9
REPENTIR	√	**remords**	√	?	√	pardon	résolution	?	√	4/9
RESPECT#1	√	√	**admiration**	**admira-tion**	**admiration**	amour	√	approbation	√	4/9
	12/18	8/18	9/18	5/18	3/18	1/18	9/18	7/18	17/18	

Tableau 2 : Ensemble des réponses

Pour ce qui est de la distribution des réponses par lexie, les meilleurs résultats reviennent aux lexies PEUR et RANCUNE : 8 réponses correctes sur 9 dans les deux cas. REGRET#I.2 et HONTE ont également eu un taux de reconnaissance acceptable : 7 et 6, respectivement. Les pires résultats reviennent aux lexies INDIGNATION et PEINE : 1 réponse correcte sur 9 dans les deux cas.

	Incorrectes			Correctes	Aucune
	QSyn	**SProche**	**"Off-mark"**		
PEUR	0	0	0	**8**	1
RANCUNE	0	0	1	**8**	0
REGRET#I.2	0	0	1	7	1
HONTE	1	0	2	6	0
REMORDS	4	0	1	4	0
RESPECT#1	3	0	2	4	0
IRRITATION	2	0	2	4	1
MÉPRIS	1	1	3	4	0
REPENTIR	1	0	2	4	2
HAINE#I.a	0	1	3	4	1
COLÈRE#1	0	2	3	4	0
HAINE#I.b	1	1	3	3	1
REGRET#I.1	0	3	3	3	0
HOSTILITÉ#2	6	1	0	2	0
HOSTILITÉ#1	3	1	3	2	0
MÉCONTENTEMENT	1	4	2	2	0
PEINE	1	4	1	**1**	2
INDIGNATION	0	2	5	**1**	1
	24	***20***	***37***	***71***	***10***

Tableau 3 : Distribution des réponses par lexie

Voici les définitions de RANCUNE et d'INDIGNATION telles que présentées aux participants du test (voir aussi la définition de PEUR, donnée précédemment). La version corrigée de la définition d'INDIGNATION peut être trouvée à la section 4.

RANCUNE (*Elle continuait à nourrir contre Florent une rancune terrible*)
~ DE L'`individu` X ENVERS L'`individu` Y À CAUSE DU `fait` Z(Y) ≡
'`Sentiment` négatif de X envers Y
causé par le `fait` suivant:
Y a fait à X un Z relativement mauvais
dont X se souvient encore et que X ne pardonne pas à Y;
ceci cause chez X le désir
que quelque chose similaire à Z arrive à Y.
Ce `sentiment` est le `sentiment` que les gens ont dans des situations semblables.'

INDIGNATION (*mon indignation contre ces crimes <ces criminels>*)
~ DE LA `personne` X CONTRE `acte/comportement` Z DE LA `personne` Y ≡
'Très fort `sentiment` négatif de X envers Z de Y
causé par le `fait` suivant:
X considère Z de Y
comme moralement ou socialement inacceptable;
ceci cause chez X le désir de faire quelque chose
pour que les Z semblables cessent d'avoir lieu.
Ce `sentiment` est le `sentiment` que les gens ont normalement dans des situations semblables.'[4]

[4] L'étiquette sémantique `personne` est une étiquette bien distincte de l'étiquette `individu` : elle désigne un ensemble d'individus ayant une fonction sociale ou politique (SYNDICAT, GOUVERNEMENT, etc.) ; cf. l'expression *personne morale*.

	QSyn	SProche
COLÈRE#1	***emportement*** ; "fam" *rogne* ; *rage* ; *fureur*	***ressentiment*** ; ***haine***
HAINE#I.a	*hostilité*#1, *animosité*	***malveillance***
HAINE#I.b	***aversion*** ; "lit" *exécration* ; *dégoût*	***intolérance***
HONTE	*déshonneur* ; *humiliation*	***gêne***
HOSTILITÉ#1	***animosité*** ; ***anthipatie***	***haine***
HOSTILITÉ#2	***guerre***	***différend***
INDIGNATION	*scandale*[5]	***réprobation*** ; ***révolte***
IRRITATION	***énervement*** ; ***colère*** ; *agacement, mécontentement*	
MÉCONTENTEMENT	***insatisfaction***, *déplaisir, irritation*	***déception*** ; ***frustration***
MÉPRIS	***dédain*** ; *irrespect* ; "lit" *mésestime*	***dégoût***
PEINE	***mal*** ; *chagrin* ; *tristesse*	***dépime*** ; ***gene***, ***malaise***
PEUR	*crainte* ; "fam" *trac* ; "pop" *trouille*	
RANCUNE	*rancoeur*, *ressentiment*	
REGRET#I.1	*douleur* ; *tristesse*	***nostalgie***
REGRET#I.2	***remords*** ; *repentir*	
REMORDS	***culpabilité*** ; ***regret***#I.2, *repentir*	
REPENTIR	***remords*** ; *regret*#I.2	
RESPECT#1	***admiration*** ; *égard, estime, consideration*	

Tableau 4 : Quasi-synonymes et sens proches des lexies visées

Les quasi-synonymes imprimés en **Arial black** proviennnent de notre corpus. Les autres quasi-synonymes (qui n'épuisent pas le répertoire de quasi-synonymes des lexies considérées) ont été ajoutés pour montrer au lecteur la richesse lexicale à laquelle on fait face (et qui pose des difficulté pour la description.)

Noter que certains **sens proches** ci-dessous ne dénotent pas des sentiments : INTOLÉRENCE 'attitude', RÉPROBATION 'attitude', DIFFÉREND 'situation' et RÉVOLTE 'action'. (En effet, une différence importante entre quasi-synonymes et sens proches est que les premiers, mais pas les seconds, ont nécessairement la même étiquette sémantique ou, du moins, la même étiquette mère).

3.2 Les résultats par participant

Comme le montre le Tableau 5 ci-dessous, le meilleur résultat revient à un non francophone expert en lexicographie explicative et combinatoire, avec 17 réponses correctes sur 18. (La seule lexie que ce participant n'a pas reconnue est PEINE, la plus « française » des lexies de notre corpus.) Le pire résultat – une réponse correcte sur 18 – a été obtenu par un francophone sans connaissances en linguistique ; l'autre participant non initié à la linguistique a eu le second résultat le moins réussi, avec seulement 3 réponses correctes.

Il existe donc, comme nous nous attendions, une forte corrélation entre le taux de réussite en reconnaissance des lexies et le niveau de connaissances en linguistique que possèdent les participants. Voici un exemple simple de la façon dont les connaissances en linguistique peuvent guider les choix dans la tâche de reconnaissance. La maîtrise de la notion d'étiquette sémantique d'une lexie L (≈ genre prochain de L) permet de ne pas confondre des lexies de sens proche, qui, comme nous l'avons indiqué plus haut, ne peuvent pas porter la même étiquette sémantique. C'est le cas, par exemple, des lexies INDIGNATION, qui est une instance de `sentiment`, et RÉVOLTE, qui, elle, est une instance de `acte`.

[5] L'acception de SCANDALE qu'on vise ici peut être illustrée par l'exemple suivant : *Il l'a épousée au grand scandale de sa famille*.

	Incorrectes			Correctes	Aucune	Francophone	Connaissances en linguistique
	QSyn	SProche	"Off-mark"				
[9]	**0**	**0**	**0**	**17**	**1**	NON	√√
[1]	1	0	5	12	0	√	√√
[3]	3	3	3	9	0	√	√
[7]	2	1	6	9	0	√	√
[2]	3	4	0	8	3	√	√
[8]	4	2	4	7	1	√	√√
[4]	5	5	2	5	1	√	√
[5]	4	4	6	3	1	√	NON
[6]	**2**	**1**	**11**	**1**	**3**	√	NON
	24	***20***	***37***	***71***	***10***		

Tableau 5 : Distribution des réponses par participant

4 Modification des définitions suite au test de reconnaissance

Nous commencerons par indiquer la définition corrigée de la lexie INDIGNATION, en expliquant comment les indications du test nous ont aidé à déceler les lacunes de la version originale de la définition, pour donner ensuite les versions finales des définitions de trois lexies quasi-synonymes de notre corpus : celles de REMORDS, REPENTIR et REGRET#I.2 (ces définitions n'ont été que légèremet modifiées par rapport à leur verison initiale).

- INDIGNATION bis

~ DE LA personne X CONTRE acte/comportement Z DE LA personne Y ≡

'Très fort sentiment négatif de $_{personne}$X envers $_{acte/comportement}$Z de $_{personne}$Y
causé par le fait suivant:
$_{personne}$X pense que $_{acte/comportement}$Z de $_{personne}$Y
viole intensément les principes de morale ou de justice de $_{personne}$X ;
ceci cause chez $_{personne}$X le désir de faire quelque chose
pour empêcher que des $_{acte/comportement}$Z semblables se reproduisent.
Ce sentiment est le sentiment que les gens ont normalement dans des situations semblables'.

Off-mark : CRIME ; JUSTICE || REGRET ; DÉGOÛT
SProche : RÉPROBATION 'attitude' ; RÉVOLTE 'action'

Mis à part le fait qu'il s'agisse de sentiments déplaisants, il est difficile de trouver un lien entre les lexies REGRET et DÉGOÛT et la lexie visée (le fait qu'on puisse regretter les choses qui provoquent l'indignation ou en être dégoûté n'est pas lexicographiquement, ou même pragmatiquement, pertinent). Le lien entre CRIME et INDIGNATION est plus clair, le crime étant une des causes possibles d'indignation et une des instances possibles de l'actant Z de la lexie correspondante. Il en va de même pour le lien entre JUSTICE et INDIGNATION, car l'actant Z de cette dernière lexie est un acte ou un comportement contraire aux principes de la justice. (Ceci veut dire que, probablement, nous aurions dû traiter CRIME et JUSTICE comme des sens proches d'INDIGNATION.)

Le lien entre INDIGNATION et RÉPROBATION s'établit par la composante 'Z(Y) est très mauvais moralement/socialement', commune aux définitions des deux lexies : ce qui indigne, à cause de son inacceptablité sociale et morale, et aussi répréhensible. Finalement, INDIGNATION et liée au RÉVOLTE par la composante 'X veut faire qqch pour empêcher Z'; cette composante reflète, justement, le caractère dynamique de l'indignation, qui est un sentiment qui pousse à l'action.[6]

La première version de notre définition d'INDIGNATION ne tenait pas suffisamment compte de ces liens et était donc trop vague.

[6] Cf. la définition d'INDIGNATION du *Petit Robert* : 'sentiment de colère que soulève une action qui heurte la conscience morale, le sentiment de la justice' → RÉVOLTE. (En passant, nous ne pensons pas qu'INDIGNATION soit une instance de COLÈRE, mais ne pouvons justifier notre point de vue ici, à cause des contraintes d'espace.)

- Série REMORDS ~ REPENTIR ~ REGRET#I.2

REMORDS (*Il avait dit au cours du procès n'éprouver aucun remords pour les attentats.*)
~ DE L'individu X À PROPOS DE L'action Y(X) ≡
'Fort sentiment négatif de $_{individu}$X causé par le fait suivant:
$_{individu}$X a/n'a pas fait $_{action}$Y, ce qui est mauvais;
$_{individu}$X est conscient du fait que $_{individu}$X n'aurait pas dû/aurait dû faire $_{action}$Y;
ceci cause chez $_{individu}$X un mal moral.
Ce sentiment est le sentiment que les gens ont normalement dans des situations semblables.'

REPENTIR (*À son réveil, son premier sentiment fut un vif repentir de sa conduite.*)
~ DE L'individu X À PROPOS DE L'action Y(X) ≡
'Sentiment négatif de $_{individu}$X causé par le fait suivant:
$_{individu}$X a/n'a pas fait $_{action}$Y, ce qui est mauvais;
$_{individu}$X est conscient du fait que $_{individu}$X n'aurait pas dû/aurait aurait dû faire Y;
ceci cause chez $_{individu}$X le désir de ne pas faire dans l'avenir une $_{action}$Y semblable
et de communiquer aux autres ce désir.
Ce sentiment est le sentiment que les gens ont normalement dans des situations semblables.

REGRET#I.2 (*Je n'ai aucun regret d'avoir fait le saut en politique*)
~ DE LA personne X À PROPOS DU fait Y ≡
'Sentiment négatif de $_{personne}$X causé par le fait suivant:
X voudrait
que $_{fait}$Y n'ait pas eu lieu ou que $_{personne}$X ait réagi différemment à propos de $_{fait}$Y
ceci est causé chez $_{personne}$X par le fait que $_{personne}$X voit les conséquences négatives de $_{fait}$Y.
Ce sentiment est le sentiment que les gens ont normalement dans des situations semblables.'

Pour clore cette section, nous aimerions proposer un outil lexicographique qui permettrait de mieux distinguer des lexies sémantiquement apparentées, en l'occurrence, les lexies dénotant des sentiments. Il s'agirait de trouver, pour chaque sentiment, une expression française qui le « représente » le mieux : quelque chose que les gens disent effectivement lorsqu'ils ressentent le sentiment en cause. Ces clichés, qui ne correspondent pas forcément à des composantes de la définition de la lexie décrite, devraient figurer dans l'article de dictionnaire de cette dernière, où ils seraient décrits au moyen d'une fonction lexicale (non standard) de type :

{ce que dit X pour exprimer son [nom de sentiment]}[7]

Voici, à titre d'illustration, les valeurs de la fonction en cause pour les trois lexies dont les définitions viennent d'être données.

REMORDS : *Comment j'ai pu ?*
REPENTIR : *Je ne le ferais jamais plus !*
REGRET#I.2 : *Si seulement je n'avais pas fait cela !*

Il existe, en outre, un cliché commun aux trois lexies (ce qui reflète bien leur nature de quasi-synonymes) : *Je n'aurais pas dû faire cela.*

5 Conclusion

La plupart des participants au test présenté ci-dessus ont eu des difficultés avec la reconnaissance des lexies à partir des définitions proposées, ce qui indique que celles-ci n'étaient pas encore adéquates en ce qui concerne leur pouvoir distinctif.

Cette situation s'explique en partie par les faits suivants :

- Les lexies dont les définitions ont été testées appartiennent à un champ sémantique difficile (cf. note 2 ci-dessus). Notons aussi que certaines de ces lexies appartiennent au vocabulaire peu courant (par ex., RANCUNE et REPENTIR) et que d'autres ont été présentées avec leur « partenaire » au sein du même

[7] Les fonctions lexicales sont des outils formels que la LEC utilise pour la description des relations lexicales, en particulier des collocation ; voir, par exemple, Wanner (ed.), 1996.

vocable (HAINE#I.a et HAINE#I.b ; REGRET#I.1 et REGRET#I.2 ; HOSTILITÉ#1 et HOSTILITÉ#2), ce qui a créé des difficultés additionnelles pour des participants non initiés à la linguistique (peu habitués aux faits de polysémie).

- Le test lui-même a été difficile, étant donné qu'on n'a offert aux participants qu'une liste des définitions à lire, sans indiquer les lexies « candidats » possibles.
- Les connaissances lexicales sont essentiellement floues, de sorte qu'on peut se demander si les gens connaissent vraiment le sens des lexies hors contexte. Elles sont aussi inconscientes, si bien qu'une rechereche consciente de l'expression pour un sens donné nécessite un certain entraînement.

À cela il faut ajouter des traits individuels des participants, notamment leur niveau du français, leurs capacités générales et leur connaissances en linguistique. (Certains participants ne seraient pas capables de reconnaître les lexies correspondantes même si on leur présentait des definitions « idéales ».)

L'exercice présenté dans cet article a permis de mettre en évidence les deux points suivants, l'un important pour le lexicographe et l'autre pour l'enseignant de langue.

- L'utilité de considérer les sens proches de la lexie L pour laquelle on veut proposer une déscription lexicographique. Une comparaison entre L et ses sens proches met en évidence les composantes saillantes du sens de L et permet ainsi un premier « dégrossissage » efficace, après quoi on peut procéder à une comparaison avec les quasi-synonymes de L, pour effectuer le *fine tuning* de la définition.
- L'utilité des connaissances SUR la linguistique pour l'acquistion des connaissances DE la langue. Il est indicatif que le meilleur score du test revient à un non francophone ayant des connaissances expertes en linguistique.

Pour ce qui est du travail à venir, il serait intéressant, d'une part, de reprendre le test de reconnaissance avec les définitions corrigées et de vrais apprenants du français langue seconde et, d'autre part, de tester l'effet de la didactisation des définitions (qui n'a été que brièvement mentionnée dans cet article) sur la reconnaissance de lexies correspondantes.

Remerciements

J'aimerais remercier Igor Mel'čuk pour ses commentaires sur une version préliminaire de cet article. Je voudrais aussi reconnaître l'aide financière octroyée au projet *Dire Autrement* par le Conseil Canadien de recherche en sciences humaines (subvention de recherche n° 410-2005-0177).

Références

Apresjan, V. & Apresjan, Ju. (1995). Metafora v semantičeskom predstavlenii èmocii. In: Apresjan (1995), *Integral'noe opisanie jazyka i sistemnaja leksikografija*. Moskva: Jazyki russkoj kul'tury, pp. 453-465.

Apresjan, Yu., Djačenko, P., Lazurski, A. & Tsinman, L. (to appear). O komp'juternom učebnike leksiki russkogo jazyka. *Russkij jazyk v naučnom osveščenii*.

Hamel, M.-J. & Milićević, J. (2007). Analyse d'erreurs lexicales d'apprenants du FLS : démarche empirique pour l'élaboration d'un dictionnaire d'apprentissage. *Revue canadienne de linguistique appliquée*, 10/1, pp. 25-45.

Iordanskaja, L. & Mel'čuk, I. (1999). Semantics of Two Emotion Verbs in Russian: BOJAT'SJA '[to] be afraid' and NADEJAT'SJA '[to] hope'. *Australian Journal of Linguistics*, 10:2, pp. 307-357.

Mel'čuk, I. (2006). Explanatory-Combinatorial Dictionary. In: Sica, G. (ed.). *Open Problems in Linguistics and Lexicography*. Monza: Polimetrica Publisher; 225-355.

Mel'čuk, I. *et al*. (1984-1988-1992-1999). *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques I-IV*. Montréal: Presses de l'Université de Montréal.

Mel'čuk, I., Clas, A. & Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Leuvin-la-neuve: Duculot.

Milićević (1997). Étiquettes sémantiques dans un dictionnaire formalisé de type « Dictionnaire explicatif et combinatoire ». Mémoire de maîtrise non publiée. Montréal : Université de Montréal.

Milićević, J. (2008). Structure de la définition lexicographique dans un dictionnaire d'apprentissage explicatif et combinatoire. In : Bernal, E. & DeCesaris, J., eds., *Proceedings of the XIII EURALEX International Congres*. Barcelona, 15-19 July, 2008. Barcelona: University Institute for Applied Linguistics, Pompeu Fabra University: 551-561.

Milićević, J. & Hamel, M.-J. (2007). Un dictionnaire de reformulation pour les apprenants du français langue seconde. Chevalier, G, Gauvin, K. & Merkle, D., eds., In : *Les apports de la sociolinguistique et de la linguistique à l'enseignement des langues en contexte plurilingue et pluridialectal*. *Revue de l'Université de Moncton*, numéro hors série 2007; 145 -167.

Polguère, A. (2003). Étiquetage sémantique des lexies dans la base de données DiCo. *Traitement Automatique des Langues* (T.A.L.), 44/2: 39-68.

Popovic, S. (2003). *Paraphrasage des liens de functions lexicales* [mémoire de maîtrise]. Département de linguistique et de traduction, Université de Montréal.

Wanner, L., ed. (1996). *Lexical Functions in Lexicography and Natural Language Processing*. Amsterdam/Philadelphia: Benjamins.

Wierzbicka, A. (1992). *Semantic, Culture and Cognition. Universal Human Concepts in Culture-Specific Configurations*. Oxford University Press.

Wierzbicka, A. (1999). *Emotions Across Languages and Cultures*. Cambrigde: Cambridge University Press.

Creating an MTT Treebank of Spanish[1]

Simon Mille, Vanesa Vidal, Alicia Burga
Barcelona Media
Av. Diagonal, 177, Planta 10
08018 Barcelona, Spain
<fname>.<lname>@upf.edu

Leo Wanner
ICREA and Universitat Pompeu Fabra
C. Roc Boronat, 138
08018 Barcelona, Spain
leo.wanner@icrea.es

Abstract

We present a cost effective strategy for the creation of a mid-size fine-grained dependency treebank of surface- and deep-syntactic structures as defined in the Meaning-Text Theory for Spanish. The strategy starts from a small seed dependency corpus, the AnCora corpus, whose annotation is considerably more coarse-grained than our target annotation. We show that this discrepancy can be bridged largely by automatic means, relying upon contextual information and leaving thus minimal work to the annotators. This allows us to develop the resources with limited human effort within a limited period of time. We also propose a preliminary evaluation of the actual amount of work that the annotation process requires.

1 Introduction

The syntactic annotation of corpora is nowadays a popular exercise in Computational Linguistics (CL). This is certainly also because the CL-community became aware that syntactically annotated corpora (or treebanks) can be used for a variety of applications – for instance, as a source for the compilation of dictionaries, as training material for machine learning-based parsing or generation, as teaching material in computer-based second language acquisition, etc.

Although the annotation with constituency trees has a longer tradition – with the first and probably still the most prominent constituency treebank being the Penn Treebank (Mitchell *et al.*, 1999) – annotation with syntactic dependency trees is gaining pace. Even more: one could say that the creation of dependency corpora is nowadays "in"; cf., for instance the Prague Dependency Treebank for Czech, containing about 96.000 sentences annotated with two levels of syntactic information – the analytical and tectogrammatical layers (Hajič *et al.*, 2006) –, the Portuguese Bosque corpus (a fully revised subset of about 9000 sentences from the Floresta corpus) (Afonso *et al.,* 2002), the Dutch Alpino treebank (van der Beek et al. 2002) with about 13350 sentences, the Swedish treebank with 11000 sentences (Nilsson et al. 2005) and the dependency version of the Penn Treebank corpus (Mitchell *et al.*, 1999).

Unfortunately, up to date, the corpus annotation initiative did not receive the due attention in the MTT community. To the best of our knowledge, so far only for Russian a stable MTT treebank is available (Apresjan *et al.*, 2006); for German, Bohnet (2003) describes experiments to map the Tiger corpus (Brants et al., 2002) onto surface-syntactic structures; and for French, the annotation of a spoken language corpus is under way, but no resources are available as yet (Gerdes, personal communication). Given the multiple prospects that a treebank offers for research and practical applications and the increased attractiveness a linguistic framework has for computational linguists in case it has to offer extended treebanks which can be used for algorithm evaluations, we think that it is very important for MTT to increase its visibility in this area.

[1] Partially funded by the Spanish Ministry of Science and Innovation (MCI FFI 2008-06479-CO2-02/FILO), in the framework of the COLOCATE project.

Our mid-term goal is the annotation of mid-size corpora (about 30.000 sentences) for Spanish, Catalan, English, German, and French, which are our primary working languages, with the structures of all levels of the MTT-model. Particular focus is put on the annotation of surface-syntactic structures (SSyntSs), deep-syntactic structures (DDSyntSs), and semantic structures (SemSs) – followed later on by the communicative structure (CommSs) at each level.[2] Starting from SSyntSs, we can automatically derive large scale Government Pattern dictionaries needed for a correct annotation with DSyntSs. Furthermore, with SSyntSs at hand, we can, at least partially, automate the process of the DSyntS-annotation and then, of the SemS-annotation.

Currently, we are working on the annotation of a Spanish corpus with SSyntSs, performing in parallel experiments on the annotation with DSyntSs and SemSs. In order to speed up the procedure of the SSyntS-annotation, we started from an existing small size Spanish dependency annotated corpus – the AnCora corpus (Martí *et al.*, 2007), whose annotation is considerably more coarse-grained than SSyntS in the Meaning-Text Theory and whose annotation conventions partially contradict the principles of MTT. But it can be semi-automatically mapped onto SSyntSs and thus serve as a seed corpus upon which the automated annotation draws.

In what follows, we describe our annotation strategy, the state of our ongoing work and our future plans. In Section 2, we provide details of the annotation procedure with SSyntSs. In Section 3 we propose a preliminary assessment of the costs of the annotation procedure. Section 4 shows how SSyntSs can be used to obtain in a relatively short time and with a relatively small effort a high quality DSyntS annotation. Section 5, finally, summarizes the paper.

2 The annotation strategy

Let us, before we delve into the details of the AnCora annotation conventions and our annotation procedure, assess the options available to annotate a corpus with SSyntSs.

2.1 Initial considerations: How to annotate a corpus with SSyntSs?

There are four alternative options for the annotation of an available (cleaned) corpus with dependency structures such as SSyntS: **I.** Manually, starting from the scratch, i.e., from a raw corpus. This option would guarantee a high quality (provided that the annotators are adequately trained and high degree of mutual agreement between the annotators is ensured), but is extremely costly. **II.** Using SSyntS-dependency parsers. Kakkonen (2006) suggests that the annotators use several dependency parsers and compare the outputs so as to produce a correctly annotated sentence. The comparison can be done automatically, based on the probability of the correctness of each parser, or manually – along with a potentially necessary correction. Unfortunately, not a single Spanish SSyntS-parser which could be used on the spot is available as yet. **III.** Starting from a constituency Treebank, mapping the constituency trees onto SSyntS dependency trees. For instance, the constituency corpus Cast3LB has already been used by Herrera *et al.* (2007) for the derivation of dependency annotations. They used the algorithm of Gelbukh *et al.* (2005) that converts constituency structures into dependency structures. Similar efforts have been made at Lund University to convert Penn-style Treebanks (Johansson and Nugues, 2007) and in the context of the ConLL shared tasks (Surdeanu *et al.*, 2008). The problem here is that it is quite difficult to obtain accurate output structures as soon as sentences are somewhat more complex. **IV.** Starting from an already existing dependency Treebank, mapping the available dependency structures onto SSyntSs. In general, given the high number of SSyntS-relations, this would imply that many SSyntS-relations will be missing and would thus need to be added either semi-automatically or manually; in addition, Spanish dependency corpora are very small. The big advantage of this option is, however, that at least the dependencies are in place.

Given the circumstances, we had to adopt the last option. The dependency Treebank from which we start is AnCora_DEP_ES (Martí *et al.*, 2007), which comprises 3512 sentences.

[2] Readers not familiar with the MTT model and terminology are asked to consult, e.g., (Mel'čuk, 1988).

2.2 Our starting point: The AnCora corpus

The AnCora dependency corpus consists of one single ConLL-format file containing 95.028 words. Figure 1 displays a sample sentence *El Gobierno de España pidió hoy al Senado que someta a votación el acuerdo*, lit. 'The government of Spain asked today to-the Senate that they-put to the vote the agreement'.

1	**El**	**el**	**d**	**da**	**gen=m|num=s**	**2**	_
2	Gobierno	Gobierno	n	np	_	5	SUJ
3	**de**	**de**	**s**	**sp**	**for=s**	**2**	_
4	España	España n	np	_		3	_
5	**pidió**	**pedir**	**v**	**vm**	**num=s|per=3|mod=i|tmp=s**	**0**	**ROOT**
6	hoy	hoy	r	rg	_	5	CC
7	**al**	**al**	**s**	**sp**	**gen=m|num=s|for=c**	**5**	**CI**
8	Senado	Senado	n	np	_	7	_
9	**que**	**que**	**c**	**cs**	**_**	**10**	**_**
10	someta	someter	v	vm	num=s|mod=s|tmp=p|per=1	5	CD
11	**a**	**a**	**s**	**sp**	**for=s**	**10**	**CREG**
12	votación	votación	n	nc	num=s|gen=f	11	_
13	**el**	**el**	**d**	**da**	**gen=m|num=s**	**14**	**_**
14	acuerdo	acuerdo	n	nc	gen=m|num=s	10	CD
15	**.**	**.**	**F**	**Fp**	**_**	**5**	**PUNC**

Fig. 1: A sample AnCora-format structure

The first column is the position of the unit in the sentence; the second, the surface form of the unit; the third, its lemma; the fourth and the fifth respectively the deep and the surface part-of-speech (POS); the sixth is an aggregation of features such as gender, number, person, depending on the POS of the unit; the seventh column is the position of the governing node, and the eigth the label of the relation with this governor.

The degree of detail and the number of the syntactic relations used in AnCora is much inferior to the set of SSynt-relations (SSyntRels): in total, 17 different labels, corresponding to about 12 of our 64 different SSynt-relations,[3] are used.[4] However, it has all syntactic dependencies marked explicitly (see below) – even if most of them are unlabelled and some of them are incorrect. In other words, each node in the annotation, except the root, has a governor. This is of great help for mapping AnCora structures onto SSyntSs. Thus, if we know that a determiner is a dependent on a noun, the relation is very likely to be *determinative*. Because this relation is not annotated in AnCora and is very frequent, being able to introduce it automatically saves a lot of time.

2.3 Annotation procedure

The annotation of a corpus with SSyntSs follows a number of basic rules which mainly originate from the notion of dependency, the characteristics of an SSyntS in MTT and considerations for further use of the SSyntS-annotated corpus:

(i) A well-formed SSyntS must be a connected tree where every node but the root must be the target of one and only one syntactical arc.

(ii) Although SSyntSs are order-free, the nodes are ordered for future machine learning applications.

(iii) The subject must be a dependent of the verbal root of a sentence structure. For instance, in *Gerard ha dejado su piso* 'Gerard has left his flat', *Gerard* is the subject of the auxiliar *ha* and not of the participle *dejado*, unlike the direct object: *Gerard* ← **subj**–*ha*–**analyt_perf**→*dejado*–**dobj**→*piso*–**det**→*su*.

(iv) Equally, the head of a relative clause is its main verb. Since an axiom of the theory is that each lexeme should correspond to one node and only one node in the tree, the relative pronoun is seen from the perspective of its function in the relative clause and not from the perspective of its

[3] See Appendix for the list of SSynt relations we use for annotation.

[4] According to the authors of the AnCora corpus, it is currently being enriched.

conjunctive properties. For instance, the phrase *Igor, que duerme* 'Igor, who sleeps' is represented as *Igor*–**relat**– **[***que*]→ *duerme* and *duerme*–**subj**→ *que*.

(v) A further consequence of the above axiom is that lexemes that occur within the same unit have to be separated. For example, *del* 'of.the' has to be split into *de+el* 'of+the', *haberlo* 'have.it' into *haber+lo* 'have'+'it', etc.

Many of these cases are not handled the same way in AnCora, which is why special attention must be paid during the SSyntS-annotation.

Another important point is that there must not be any ambiguity of the valency patterns in the SSyntS so as to facilitate the annotation at the more abstract levels DSyntS and Sem and derivation of a Government Pattern (GP) dictionary. To ensure this, we introduce several SSyntRels for the same grammatical function but different underlying GPs; for instance, *obl_obj1*/*2*/*3* for indirect objects (with the index marking the corresponding semantic actant slot). This allows us to obtain GPs by retrieving the corresponding DSyntS information without any ambiguity, and then (partially) derive the DSyntSs from the SSyntSs. However, given that for other purposes such a fine-grained and semantically oriented annotation is not appropriate (e.g., for training of a syntactic dependency parser), we maintain in parallel a version of the annotation in which only purely syntactic relations appear, i.e., in which *obl_obj1*/*2*/*3* relations are merged into the single relation *obl_obj*.

The annotation procedure that draws upon the above rules comprises several stages:

(1) Automatic projection of the annotations of the 3.512 sentences from AnCora onto rudimentary SSynt-like structures. This stage consists of two substages:
   (1a) A simple script maps in a one-to-one fashion AnCora relations/features onto SSynt-like relations/features.
   (1b) Using the graph transduction workbench MATE (Bohnet et al., 2000; Bohnet, 2006), inference rules derive from the topology of the AnCora structure additional SSynt-relations that are not available in AnCora.

(2) Manual revision of the structures obtained in Stage 1 by a team of grammarians, who follow detailed guidelines. For the revision work, MATE's graph editor is used.

(3) Training of a machine learning-based dependency parser (Bohnet, 2009) with the obtained SSyntSs and its application onto a new subcorpus of about 3.000 sentences.

(4) Manual revision of the structures obtained in Stage 3 and extension of the parser training corpus by these structures (cf. Hwa (2001) for more details and references on this particular method);

(5) Repetition of Stages 3 and 4 until the SSyntS annotated corpus reached the desired size. With each iteration, the quality of the parsing results improves, such that the cost of the manual revision decreases considerably.

During Stage (1a), the goal is thus to simply convert all labels – attribute/value pairs and arcs – into labels used in MTT's SSyntSs. For instance, the subject relation "SUJ" becomes "subj", the direct object relation "CD" becomes "dobj", the determinative POS feature "d" becomes the feature/value pair "spos=determiner" and so on. To facilitate higher quality parsing, we also introduce the POS tags from the Penn Treebank set (Mitchell *et al*., 1993). A simple script handles such one-to-one correspondences and provides intermediate ConLL-structures with appropriate tags (not shown here because too similar to Fig. 1; cf. Fig. 2 for graphical representation). The modified AnCora structure is imported into MATE's graph editor, where all dependency relations and the precedence relations (relations "b") as available in the ConLL structure can be visualized; cf. Fig. 2.

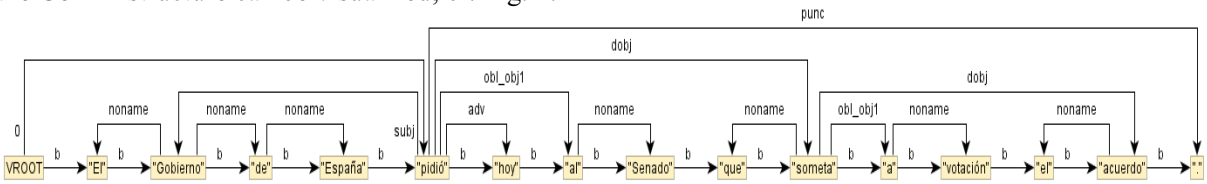


Fig. 2: Graphical representation of an AnCora Structure converted into a preliminary SSyntS

All empty relation names are mapped to “noname” labels. We see that there are quite a few arcs that are not labelled; that *al* ‘at.the’ is one single node; that some labels are wrong (thus, “obl_obj1” stands for actant 2 and here both dependents of “obl_obj1” are, in fact, actant 3 of their governing verb); and that some dependencies are erroneous – as, e.g., the conjunction *que*, which should govern the main verb of the subordinated clause and not be a dependent of it.

The second mapping (Stage 1b), performed automatically by using a small graph-transformation grammar of 55 rules in the MATE workbench, corrects some of these errors. Most of the rules simply check in the AnCora structure the nature of two nodes linked by arcs labelled “noname” and introduce an SSyntRel. Consider for instance the rule that introduces the *appos*(itive) relation:

```
?Xl {   dpos=N                                  rc:?Xr {<=> ?Xl
        noname→?Yl {dpos=N}        ➔            appos→ rc:?Yr {<=> ?Yl}}
        b→ ?Yl}
```

This rule states that if two nodes ?Xl and ?Yl that have the same deep part-of-speech N are linked by an arc “noname”, and if ?Yl follows ?Xl, then an arc *appos* is added from ?Xl to ?Yl in the target structure (the ‘rc:’ prefix in the right hand side of the rule is due to internal MATE codification conventions and can be ignored here). Other types of rules handle the separation of nodes or check the root of a verb group. Applied to the structure in Fig. 2, the transformation grammar gives us the following structure in Fig. 3:

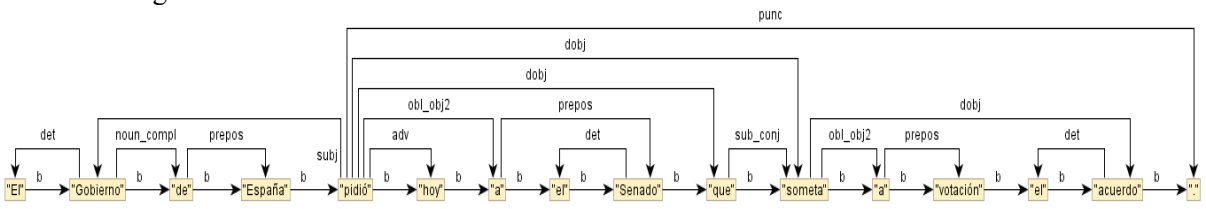


Fig. 3: Structure after stage (1b)

The *al* node is now split into a preposition node and a determiner node; all relations are added; all arcs are labelled with a dependency relation; the “obl_obj1” labels have been mapped to “obl_obj2”, which corresponds to the actant 3 slot of the governor; the dependency that we expect between the conjunction *que* and the verb *someta* ‘put’ is now visible, hence the relation “dobj” between the governor *pidió* and this conjunction. However, it can be also observed that the automatic mapping introduces new errors into the annotation, such as multiple edges, which must be corrected manually in Stage 2.

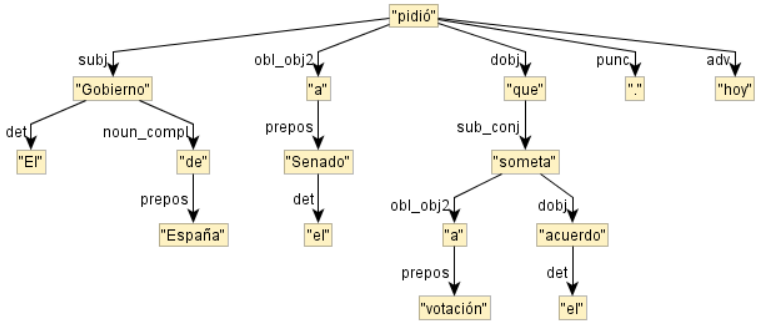


Fig. 4: Correct SSynt dependency tree

For such simple structures as the one in the figures above, already the first mapping is very efficient and the manual corrections can thus be kept to the minimum. In this particular case, just one arc has to be

removed to get the final structure (shown without the precedence relations in Fig. 4 above).[5] When the structure is more complex, there are, of course, more errors, be it in the original corpus or during the second mapping in Stage 1b. The main errors for each level are detailed in the next section.

Stage 2 is carried out by a team of annotators trained in MTT. In order to ensure a high quality annotation, structures annotated by one annotator are verified by two other annotators.

Once the 3512 sentences from the original AnCora corpus have been annotated with SSyntSs, the training of the machine learning based parser starts. The training algorithm implemented by B. Bohnet (2009) delivers models for a parser that reached an accuracy of about 96% for German (with respect to both dependency links and labels); we are confident that we reach a similar accuracy for Spanish.

Stage 2 has recently been completed. In order to ensure a high quality annotation, structures annotated by one annotator are currently cross-checked by two other annotators. We expect this procedure to be finished by September 2009.

3 Assessment of the annotation

In order to be able to assess the costs of the annotation of a corpus with such detailed dependency information as SSyntSs, it is essential to be aware of the errors encountered at the different stages of the annotation procedure as well as of the manual workload envisaged by the annotators.

3.1 Error evaluation

During Stage (1b) of our annotation procedure, two main types of errors that directly influence the manual workload of the annotators are introduced: (i) wrong choice of actants, especially for nouns, and (ii) over-generation of arcs. Errors of type (i) arise because it is impossible to know from the syntactic structure to which semantic argument a syntactic actant corresponds. In Spanish NPs, actants of a governing noun are related to it by the preposition *de*: *lista de paro*, *presidente de+l gobierno*, etc. Therefore, in the AnCora-SSyntS mapping, only one rule introduces nominal actantial relations. By default this is the relation that corresponds to the first actant, i.e. nominal completive – as, e.g., encountered in *una lista de escuelas* 'a list of schools'. However, in many cases, it is actually the second actant (as in *el presidente de Francia* 'the president of France'), or even a third or a fourth, or an attribute, i.e., not an actant at all (as in *mesa de madera* 'wooden table'). That is, the annotator must pay close attention to this phenomenon in particular.

Errors of type (ii) are due to the fact that the application of the mapping rules is not sufficiently constrained. Indeed, the rules are preferred to apply even in uncertain cases in order to avoid that they miss some relations: for the annotator, it is easier and faster to remove an arc than to add a new one.

Although the mapping during Stage (1b) introduces errors, it also corrects suboptimal (from the point of view of SSyntSs) choices made by the AnCora annotators, such as leaving many dependency arcs unlabeled (see footnote 4 above) or treating some word combinations as single units, even if they are not at the syntactic level, as shown *supra*.

Several annotation characteristics of the AnCora corpus also required massive manual intervention on our part because they could not always be handled by mapping grammar rules. The most significant of them are:

- an adjective positioned before a noun is considered the head of the adjectival phrase a part of which is the noun; accordingly, the adjective is considered governor of various dependents of the noun: in MTT, the noun is the governor of its adjectival modifiers;
- non-finite verbal heads in auxiliary constructions and raising/control constructions are considered to be syntactic heads of verb groups: although it is the *semantic* head of the group, in syntax, the finite verb has to be the governor;
- the coordinate constructions have been partially left aside;

[5] For annotators' convenience, the "linearized" trees can be shown in the tree format – for instance, to facilitate the connectivity check.

- the internal dependencies in relative clauses are often missing.

So, knowing this, what is the amount of work that an annotator has to invest in order to carry out his/her task? Let us assess this in the next subsection.

3.2 Extent of the manual workload

In order to carry out a preliminary evaluation on the manual workload, we picked randomly 50 sentences out of our annotated set and manually counted the modifications that had been performed by the annotator. Three types of manipulations have been identified, by order of descending complexity: (1) create nodes (includes creating and labelling arcs); (2) create or move an arc (includes labelling the arc); (3) label an arc that is correctly positioned.

We counted 9 interventions of type 1, 366 of type 2, and 77 of type 3. This gives an average of about 0.2 creations of nodes, 7.3 creations of arcs, and 1.5 arc re-labellings per annotated sentence. While these figures seem low compared to what is usually needed to annotate sentences, it should not be forgotten that what takes more time is not editing a graph, but elaborating all the dependencies between the units of the sentence. The process is certainly made much easier, but the workload remains important.

4 How to obtain a DSyntS annotation?

As already mentioned, the richness of the SSynt dependencies makes the SSyntS very informative. In this section, we show that it grants direct access to DSyntSs.

As illustrated in the previous sections, our SSynt annotation foresees different names for the same syntactic dependency, depending on the valency slot occupied by the dependent. In the SSyntS of Fig. 4, for instance, we find four predicates, *gobierno* 'government', *pedir* 'ask', *someter* 'put'[6] and *acuerdo* 'agreement'. What can be deduced from Fig. 4 is that:

- *gobierno* has an actant 1 (realized here by the SSyntRel 'noun completive');
- *pedir* has an actant 1 ('subjectival'), an actant 2 ('direct objectival'), and an actant 3 ('oblique objectival 2');
- *someter* has an actant 2 ('direct objectival'), and an actant 3 ('oblique objectival 2'); the first actant does not have to be realized;
- *acuerdo* appears without any actant realized. We cannot draw any conclusion with respect to its actant structure.

A simple mapping grammar in MATE extracts this lexical information from the SSyntS in Fig. 4 in terms of the following lists of attributes, corresponding to the "syntactic combinatorial zone" as described in (Mel'čuk*,* 2006):

– *gobierno* { **dpos**=N
I_dpos=N **I_spos**=proper_noun **I_rel**=noun_compl **I_prep**=de }
– *pedir* { **dpos**=V
I_dpos=N **I_spos**=proper_noun **I_rel**=subj
II_dpos=V **II_spos**=verb **II_rel**=dobj **II_prep**="que" **II_mood**=SUBJ **III_dpos**=N **III_spos**=proper_noun **III_rel**=obl_obj2 **III_prep**="a" }
– *someter* { **dpos**=V
II_dpos=N **II_spos**=noun **II_rel**=dobj
III_dpos=N **III_spos**=noun **III_rel**=obl_obj2 **III_prep**="a" }
– *acuerdo* { **dpos**=N **ssynt_actant**=NO }

Pedir 'ask', for instance, contains four lines of attribute/value pairs: in the first line appears its deep part-of-speech (*dpos*); the second line presents the information corresponding to its first DSynt actant: this actant is a proper noun, linked by the relation "subj" to its governor; in the third line, the information about the second DSynt actant is stored: it is a verb linked to *pedir* by a direct objectival relation 'dobj',

[6] *Someter* cannot not generally be translated by 'put'; here, it is, actually, the value of a lexical function; see below.

such that this verb is introduced by *que* 'that' and is in the subjunctive mood. Similarly, the last line describes the third actant of *pedir*.

Of course this is not the only way to use the verb *pedir* in Spanish. First, there are other senses corresponding to *pedir*, such as 'order' (a coffee) or 'beg' (for money). These cases are left aside for the moment since those instances of *pedir* are different lexical units, and thus also different entries in the dictionary. Second, for the same lexical unit meaning 'ask' – let us call it *pedir_1* – there are several GPs that can be realized; consider the following sentences that exemplify a variety of partial GP instantiations:[7]

(1) *El jefe le pidió a Elena que escribiera ese informe.* 'The boss asked *PREP* Elena that [she] wrote the report': actant II is a subjunctive verb with governed conjunction;

(2) *El jefe le pidió a Elena escribir ese informe.* 'The boss asked *PREP* Elena [to] write the report': actant II is an infinitive verb without preposition;

(3) *Le pidió un favor a Elena.* '[He/she] asked [for] a favour to Elena': actant II is a noun; no actant I is visible.

All GPs will eventually appear in the entry for *pedir_1* in the lexicon; how this is achieved is beyond the scope of this paper. What matters here is that any GP of any lexical unit can be stored in the dictionary, with all properties of the governed element that are required by the governor (POS, mood, finiteness, etc.), and so on.

With such a dictionary at hand, it is very easy to derive DSyntS since one of the main challenges of the SSynt-DSynt transition is to distinguish semantic prepositions from syntactic (*governed*) prepositions; the latter being the only ones stored in the dictionary. For example, in *Marc le pidió a Elena que le llamase por teléfono*, lit. 'Marc [her] asked to Elena that him she.calls by [the] phone', the meaningless governed preposition *a* 'to' does not appear in the DSyntS (neither does *que* 'that'), whereas the preposition *por* 'by', which has a meaning, namely the way how Marc asked Elena to call, has to appear as a node label in the DSyntS.

In the case of the SSyntS in Fig. 4, using such a dictionary, we can readily derive a DSyntS shown at the left hand side of Fig. 5. This DSyntS is "nearly" correct. It is not entirely correct because it does not take into account the notion of *lexical function*, LF (Mel'čuk, 1996). Thus, the verb *someter* is, in fact, the value of an LF, namely CausOper_2 applied to the keyword *votación*; cf. the right hand side DSyntS in Fig. 5.[8] In other words, in order to annotate DSyntSs appropriately, LFs have to be introduced directly by the annotator;[9] however, the total amount of work necessary for the compilation of a DSyntSs corpus remains rather low once the SSyntSs corpus has been built.

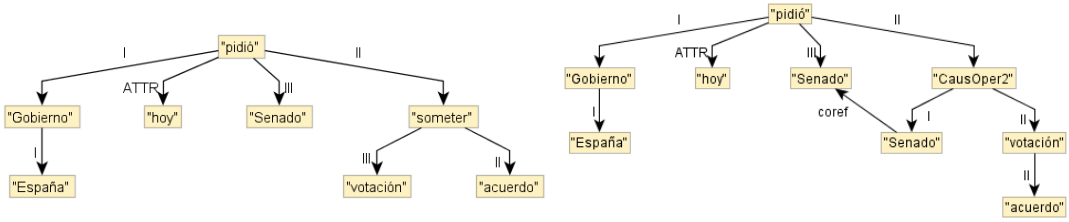


Fig. 5: Automatically derived (left hand side) and correct (right hand side) DSyntS corresponding to SSyntS in Fig. 4

Once the DSyntS annotation is in place, we can approach the SemS annotation, which will be perforce a simplified SemS, without lexical decomposition.

---

[7] Only actant II is detailed since the first and the third are the same in these instances of this lexical unit.

[8] Due to the limitations of the graphical editor, the theoretically bidirectional coreference relation 'coref' between the two nodes of "Senado" is shown as uni-directional.

[9] The work on the automatic recognition of LFs in corpora as discussed, e.g., in (Wanner et al., 2006) is still too preliminary to be used for high quality annotation.

5 Conclusions and Future Work

We presented a cost effective strategy for the creation of a mid-size fine-grained dependency treebank of MTT's SSyntSs for Spanish. The strategy draws upon a small seed dependency corpus, the AnCora corpus, whose annotation is considerably more coarse-grained than our target annotation. We have shown that this discrepancy can be bridged largely by automatic means, relying upon contextual information and leaving thus minimal work to the annotators. This facilitates the development of resources with limited human effort, within a limited period of time. The availability of the SSynt treebank will allow us to pursue research in a number of different directions. For instance, once annotations of several layers are available, we can use machine learning techniques for automatic learning of sentence generation or analysis grammars.

Acknowledgements: We are grateful to Antònia Martí for making the AnCora corpus available to us. Many thanks also to Igor Mel'čuk for his generous help with respect to all kinds of problems related to dependency and for his helpful comments on this paper, to our colleagues in the TALN group, especially Gabriela Ferraro, for their support, and to the two anonymous reviewers for their helpful comments. All remaining errors and omissions are, as always, our full responsibility.

References

Afonso, S., Bick, E., Haber, R., and Santos, D. (2002). "Floresta sintá(c)tica": A treebank for Portuguese, in M. González Rodríguez and C. Paz Suárez Araujo (eds.), In *Proceedings of LREC*, 29-31. Las Palmas de Gran Canaria, Spain. ELRA, pp.1698-1703.

Apresjan, Ju., Boguslavsky, I, Iomdin, B., Iomdin, L., Sannikov, A., and Sizov. V. (2006). A Syntactically and Semantically Tagged Corpus of Russian: State of the Art and Prospects. In *Proceedings of LREC*, 1378-1381. Genova, Italy.

Beek van der, L., G. Bouma, R. Malouf, and G. van Noord. (2002). "The Alpino dependency treebank". In *Linguistics and Computers. Selected Papers from the 12th CLIN Meeting*. Twente, The Netherlands, 8-22

Bohnet, B., A. Langjahr and L. Wanner. (2000). "A Development Environment for an MTT-Based Sentence Generator". *Proceedings of the First International Conference on Natural Language Generation*. Mitzpe Ramon, Israel, 260-263

Bohnet, B. (2003). "Mapping Phrase Structures to Dependency Structures in the Case of Free Word Order Languages". In *Proceedings of MTT conference 2003*, Paris

Bohnet, B. (2006). *Textgenerierung durch Transduktion linguistischer Strukturen*. DISKI 298. Akademische V. G., Berlin.

Bohnet, B., (2009). "Synchronous Parsing of Syntactic and Semantic Structures". To appear in *Proceedings of the Fourth International Conference on Meaning-Text Theory*, Montreal.

Brants, S.; Dipper, S.; Hansen, S.; Lezius, W. and Smith, G. (2002). "The TIGER Treebank". In *Proceedings of the Workshop on Treebanks and Linguistic Theories*. Sozopol.

Gelbukh, A., Torres, S. and Calvo, H. (2005). "Transforming a Constituency Treebank into a Dependency Treebank". In *Proceedings of the X Conference of the Spanish Association for Artificial Intelligence*.

Hajič, J. et al. (2006). *Prague Dependency Treebank 2.0* . In Linguistic Data Consortium, Philadelphia.

Herrera, J., *et al* (2007). "Building Corpora for the Development of a Dependency Parser for Spanish Using Maltparser". In *Procesamiento del Lenguaje Natural*, nº39, pp. 181-186. Spain.

Hwa, R. (2001). On minimizing training corpus for parser acquisition. In *Proceedings of the Fifth Computational Natural Language Learning Workshop*, Toulouse, France, July.

Johansson, R. and Pierre Nugues (2007). "Extended Constituent-to-dependency Conversion for English." In *Proceedings of NODALIDA 2007*. Tartu, Estonia.

Kakkonen, T. (2006). DepAnn - An Annotation Tool for Dependency Treebanks. In *Proceedings of the 11th ESSLLI Student Session at the 18th European Summer School in Logic, Language and Information*, pp. 214–225. Malaga.

Martí, M.A., Taulé, M., Márquez, L., Bertran, M. (2007): "Ancora: A Multilingual and Multilevel Annotated Corpus", http://clic.ub.edu/ancora/publications/

Mel'čuk, I.A. (1988). *Dependency Syntax: Theory and Practice*, Albany, N.Y.: The SUNY Press.

Mel'čuk, I.A. (1996). Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon. In L. Wanner (ed.) *Lexical Functions in Lexicography and Natural Language Processing*. Amsterdam: Benjamins.

Mel'čuk, I.A. (2003). Levels of Dependency in Linguistic Description: Concepts and Problems. In V. Agel, L. Eichinnger, H.-W. Eroms, P. Hellwig, H. J. Herringer, H. Lobin (eds): *Dependency and Valency. An International Handbook of Contemporary Research,* vol. 1, Berlin - New York, W. de Gruyter, 188-229

Mel'čuk, I.A. (2006). Explanatory Combinatorial Dictionary. In G. Sica (ed.). Open Problems in Linguistics and Lexicography. Monza, Italy: Polimetrica, 225-355.

Mitchell P. M., B. Santorini, and M.A. Marcinkiewicz (1993). "Building a Large Annotated Corpus of English: The Penn Treebank", In *Computational Linguistics*, 19(2):313– 330.

Mitchell P.M., B. Santorini, M.A Marcinkiewicz, and A. Taylor (1999). "Treebank-3", LDC, Philadelphia.

Nilsson, J., J. Hall and J. Nivre. (2005). "MAMBA Meets TIGER: Reconstructing a Swedish Treebank from Antiquity". In *Proceedings of NODALIDA 2005 Special Session on Treebanks for Spoken and Discourse*, Copenhagen Studies in Language 32, Joensuu, Finland, pp. 119-132.

Surdeanu, M., R. Johansson, A. Meyers, L. Màrquez, and J. Nivre (2008). "The CoNLL-2008 Shared Task on Joint Parsing of Syntactic and Semantic Dependencies." In *Proceedings of the 12th Conference on Computational Natural Language Learning* (CoNLL).

Wanner, L., B. Bohnet, M. Giereth, and V. Vidal. (2006). "The first steps towards the automatic compilation of specialized collocation dictionaries". In *Terminology*, 11(1):143-180.

Appendix: List of the 64 SSyntRels used for the annotation (inspired by Mel'čuk 2003).

For parser training: ♣=SSyntRel is merged with *adv*; ♦=SSyntRel is merged with *obl_obj*; consecutive♠=SSyntRels are merged together.

adjunct (adjunctive): Vale/Juan,<-*adjunct*- vamos; Pero, <-*adjunct*-[no lo]-sabían; por <-*adjunct*-[ejemplo,]-considera [esta solución]; [Mañana, vendrá] al_menos <- *adjunct* –Pedro;
adv (adverbial): etiquetar-*adv*->rápidamente; volvió-[el]-*adv*->día [siguiente]; volver-*adv*->corriendo; aproximadamente <-*adv*-veinte; [sabe] cuándo<-*adv*-viene;
adv_abs (absolutive adverbial): Terminada <-*adv_abs*-[la guerra]-,-volvieron [a casa]; el pan<-a*dv_abs*-[en la mano]-,-salió; vi a Vasco-,-*adv_abs*-> guitarra [en mano];
adv_clitic (clitic adverbial): producir-*adv_clitic*->le [otra nidada]; le <-*adv_clitic*-plantó [un árbol];
adv_mod (modificative adverbial): [Muy] tranquilo,<- *adv_mod*-viajaba [a menudo];
adv_obj1♣ (objectival adverbial 1): viene-*adv_obj1*->aquí; va- *adv_obj1*-> adentro; [me] siento- *adv_obj1*->bien;
adv_obj2♣ (objectival adverbial 2): [Lo he] traído-*adv_obj2*->aquí; [lo] calificó-*adv_obj2*-> positivamente;
agent (agentive): escrito-*agent*->por [Leo];
analyt_fut (future analytical) : va-*analyt_fut*-> a conducir;
analyt_pass (passive analytical) : es-*analyt_pass*-> conducido;
analyt_perf (perfect analytical): ha-*analyt_perf*-> conducido;
analyt_progr (progressive analytical): está-*analyt_progr*-> conduciendo;

appos (appositive): [el] presidente-*appos*-> Obama; [la] nebulosa-*appos*-> de [Orion];
appos_descr (descriptive appositive): [el] presidente,-*appos_descr*->Obama, […];
attr (attributive): casa-*attr*->sin [ventanas]; mesa-*attr*-> de [madera]; niño-*attr*-> con [gafas];
attr_descr (descriptive attributive): [el profesor] Wanner,-*attr_descr*->de [Barcelona, estuvo aquí];
aux_phras (phraseological auxiliary): lo_más-[claro]-*aux_phras*-> posible; tomar-*aux_phras*-> en [cuenta];
aux_refl (reflexive auxiliary): me<-*aux_refl*-afeito; se<-*aux_refl*-miran; se<-*aux_refl*-come [un conejo];
bin_junct (binary junctive): o<-[Barça]-*bin_junct*-o [Real]; desde-[dos]-*bin_junct*->hasta [cuatro];
compar (comparative): mejor-*compar*->que, tan-[bonito]-*compar*->como [alto/Juan];
compar_conj (comparative conjunctional) : [mejor] que-*compar_conj*->tú;
compl1♠ (completive 1): [La frase] resulta-*compl1*-> buena;
compl2♠ (completive 2): [Vane] encuentra-[la semántica]-*compl2*-> fácil;
compl_adnom (adnominal completive): Los-*compl_adnom*-> de [la ciudad];
coord (coordinative): sentido-*coord*->y [texto], dependencia-*coord*->o [constituyente];
coord_conj (coordinate conjunctional): [sentido] y-*coord_conj*-> texto;
copul (copulative): Igor es-*copul*->guapo/[un] hombre;
copul_clitic (clitic copulative): [Igor] lo<-*copul_clitic*-es;
det (determinative): este<-*det*-artículo; un<-*det*-gato; lo<-*det*-divertido [es que caí];
dobj (direct objectival): [Igor] come-*dobj*->gatitos; quiere-*dobj*->que [vengas]; [Leo] ve-*dobj*-> a [Marga];
dobj_clitic (direct objectival clitic): haz-*dobj_clitic*->lo; la <-*dobj_clitic*-mira;
dobj_quot (quotative direct objectival): ha gritado-*dobj_quot*-["¡]->Gooool!";
elect (elective): [el] mejor-*elect*->de [los pintores], [el] más-[tonto]-*elect*->en [Barcelona];
inf_obj1♠ (infinitival objectival 1): [Juan] piensa-*inf_obj1*->ganar; [su]deseo-*inf_obj1*->de [venir];
inf_obj2♠ (infinitival objectival 2): [lo] empuja-*inf_obj2*-> a [venir];
juxtapos (juxtapositive): Es-[muy potente:]-*juxtapos*-> puede [destruir un país en 10 minutos];
modal (modal verb): [Manuel] puede-*modal*-> venir (Closed list: poder, deber, querer);
modif (modificative): gato-*modif*->pelado; pequeño<-*modif*-árbol ; [un] chico-*modif*->más;
modif_abs (absolutive modificative): [los] gatos-,-*modif_abs*-> incluidos [los negros, me gustan];
modif_descr (descriptive modificative): [las] ventanas,-*modif_descr*->sucias [y rotas, se caen];
noun_compl♦ (noun completive): [las] gafas-*noun_compl*-> de [Pep]; [una] traducción-*noun_compl*-> de [Stefan]; falta-*noun_compl*-> de [chocolate]; aceite-*noun_compl*->de [oliva];
num_junct (numeral junctive): treinta-*num_junct*->y [tres]; tres<-*num_junct*- mil (3000);
obj_copred (object copredicative): [Igor] quiere-[la estructura]-*obj_copred*-> conectada;
obl_obj1♠ (first oblique object): ir-*obl_obj1*-> a [la playa]; tener-*obl_obj1*-> que [comprar]; presidente-*obl_obj1*-> de [Francia]; traducción- *obl_obj1*-> de [este texto]; gracias-*obl_obj1*-> a [Leo]; capaz-*obl_obj1*-> de [hablar];
obl_obj2♠ (second oblique object): vender-[un disco]-*obl_obj2*->a [Marc por 10 €]; suplemento-*obl_obj2*-> de [economía de la Vanguardia];
obl_obj3♠ (third oblique object): vender-[un disco a Marc]-*obl_obj3*-> por [10 €];
obl_obj_clitic1♠ (first oblique object clitic): [la virgen se] le <-*obl_obj_clitic1*-aparece [cada miércoles];
obl_obj_clitic2♠ (second oblique object clitic): dar-*obl_obj_clitic2*->le; le <-*obl_obj_clitic2*-da;
prepos (prepositional): en-*prepos*-> cama;
prolep (prolepsis): Yo<-*prolep*-,- [lo que veo]-es [una torre];
quant (quantitative) : [tres] mil <-*quant*-personas;
quasi_coord (quasi coordinative): a[l]-[norte]-,- *quasi_coord*-> allí [donde casi nadie mira];
quasi_subj (quasi subjectival): Eso<-*subj*-sí-,- *quasi_subj*-> que [venga mañana];
relat (relative): [el] gato-[que]-*relat*-> está [aplastado]; [el] edificio-[en que]-*relat*-> trabajamos;
relat_descr (descriptive relative): [este] artículo,-[que]-*relat_descr*->leí [ayer, es corto]
relat_expl (explicative relative): Juan salta-,-[lo_que]-*relat_expl*->sorprende [a su madre];
restr (restrictive): más <-*restr*-frecuente; no <-*restr*-bebe; sólo <-*restr*-fuma;
sequent (sequential): cuarenta-*sequent*-> hasta [cincuenta]; [el partido] Real-*sequent*-> Barça [terminó 2-6];
sub_conj (subordinate conjunctional): [es verdad] que-*sub_conj*-> [ayer llovió];
subj (subjectival): [el] sol <-*subj*-baja; parece-[imposible]-*subj*-> ver [este detalle]; es-[a Verónica]-*subj*-> que [le he escrito]; claro/sí- *subj*-> que [te llamaré];
subj_copred (subject copredicative): [Mariano] volvió-*subj_copred*-> rico;
subj_quot (quotative subject): 'Algo' <-*subj_quot*-es [sujeto de esta frase];
+ (only for parser training) **punc** (punctuation)/**punc_init** (initial punctuation)

Mode of Action Nouns and Their Diatheses

Elena Paducheva
Moscow, Russian Academy of Science
elena708@gmail.com

Abstract

Conceptual apparatus of "Meaning – Text" theory suggests a thoroughly elaborated classification of predicative nouns. This classification distinguishes nouns of action, nouns of subject, object and second object, nouns of instrument, place, and other circumstances, see Mel'čuk 1974: 87, Apresjan 1974: 48, 199. This paper deals with Russian mode of action nouns (such as *počerk* 'handwriting', *poxodka* 'step'). It is demonstrated that when a mode of action noun is motivated by a transitive verb it has a special diathesis in which the subject can be expressed not by instrumental case, as usual, but by a possessive.

1 Introduction

Possessives in Russian can be divided into two groups.

1) Possessive pronouns – personal, such as *moj* 'my' *tvoj*, 'your', *vaš* 'your', *naš* 'our' and, with some reservations also 3d person pronouns *ego* 'his', *ee* 'her', *ix* 'their'; reflexive pronoun *svoj*; relative pronoun *čej* 'whose'; indefinite pronouns *čej-to*, *čej-nibud'*, *čej-libo*, *koe-čej*; and negative *ničej*.

2) Possessive adjectives in *-ov*, *-in*, *-ovskij*. Adjectives ending in *-ij*: *lisa–lisij* 'fox–fox's', are drastically different and should be studied separately.

Both groups of possessives are described at length in (Shmelev 2008) devoted to possessives in the context of concrete nouns (predmetnye imena). I am interested in possessives in the context of predicative nouns, in particular, in the context of deverbal nouns (such as *interpretacija* 'interpretation' from *interpretirovat'* 'interpret') and verbal nouns, such as *koncepcija* 'conception' from a hypothetical verb meaning 'create a conception'.

2 Possessive pronouns

In its basic meaning a possessive pronoun expresses the idea of belonging to a person. In other words, it denotes a person as a possessor. As is known, the idea of possession appears only in the prototypical context – in the context of a name of an object (*moja kružka* 'my cup', *ee rasčeska* 'her comb'). In the context of a predicate noun the possessive pronoun does not express possession and does not necessarily refer to a person. In the first place this concerns the 3d person pronouns: *ee otsutstvie* can contextually mean 'lack of *water*', 'lack of *freedom*', etc.

As regards referential status of possessive personal pronouns, they are all alike in allowing only referential use. In fact, a non-possessive personal pronoun may refer to a generic NP, while possessive pronouns cannot be used this way:

(1) Surovyj Dant ne preziral *soneta*. V *nem* [sonete] žar ljubvi Petrarka izlival. 'Severe Dante didn't disdain *a sonnet*. Petrarch expressed the flame of his love in *it*'

3 Possessive adjectives

Now about possessive adjectives ending in *-ov* (*-ev*), *-in* (*nin*) and *-ovskij*. They also refer to a concrete person; we have all rights to say that possessive adjective is a POSSESSIVE FORM of a referential noun – very often it is a proper name.

There is a large and ever growing class of possessive adjectives formed with the help of the suffix *-ovskij*: *vendlerovskaja klassifikacija* 'Vendler's classification' *rixterovskoe ispolnenie* 'Richter's presentation' etc. While the pattern with *-ov* is non-productive – E. A. Zemskaja (1992: 76) gives only one example *Gulliverov* 'Gulliver's' – adjectives ending in *-ovskij* with possessive meaning are easily formed, practically without limitations, from foreign names recognizable as belonging to the 2d declension. The same can be said about adjectives that can be formed with the help of the suffix *-in* from names of the first declension (*Vanin* 'Vanja's', *Šarlottin* 'Sharlotta's').

Adjectives ending in *-ovskij*, as well as possessive pronouns, may have both possessive function and the function of the subject. Thus, we have full right to speak about common possessive semantics of pronouns and adjectives.

In so far as a noun motivating a possessive is referential it is possible to transfer a possessive construction to a genitive one (the example below, as well as most of the examples that follow are taken from National corpus of Russian, site in the Internet www.ruscorpora.ru):

(2) Inogda *sviridovskie vzgljady na grjaduščee* prosto donel'zja černy i mračny = *vzgljady Sviridova na grjaduščee* [Sviridov's views on the future = views of Sviridov on the future]

4 Mode of action nouns and their possessive-genitive diathesis

Now let us return to predicative mode of action nouns motivated by a transitive verb. Remarkably, they constitute a context in which possessive pronouns and adjectives effectively oust the genitive in its subject expressing function. See an example (on the basis of Padučeva 1984, 1974: 203):

(3) a. *tvoe* ispolnenie Šopena '*your* rendition of Shopen'
b. *Rixterovskoe* ispolnenie Šopena '*Richter's* rendition of Shopen'
c. *ispolnenie *Rixtera* [gen] Šopena [gen]

As follows from (3c), in the context of the genitive object, the subject cannot be expressed by the genitive (though two genitives are not excluded in principle, cf., e.g., acceptable *lišenie brata nasledstva* 'brother's deprivation of heritage'); but it can be expressed by a possessive pronoun, as in (3a), and a possessive adjective, as in (3b).

The problem is, how to delimitate the class of predicative nouns allowing for the possessive construction illustrated by examples (3a) and (3b).

As is known, the participant subject of a predicative noun is usually expressed by Instrumental, the genitive being reserved for the object:

(4) razrušenie Novgoroda moskovskim *knjazem* 'destruction of Novgorod by the Moscow prince'

Still there are contexts in which only genitive can be used to express the subject. For example, the subject is expressed by the genitive if the substantive is motivated by an intransitive verb:

(5) priezd *brata* 'arrival *of the brother*'

(6) vozvraščenie *Ivana* na rodinu 'return of Ivan to motherland'

If the substantive is motivated by a transitive verb its subject can be expressed by a genitive only in some special contexts.

1. The genitive subject is possible in the context of nouns derived from verbs *ljubit'*, *uvažat'*, which, being transitive, do not license genitive object in the corresponding derived nouns (*ljubit' apel'siny* 'love oranges' gives *ljubov' k apel'sinam* 'love to oranges' and not **ljubov' apel'sinov* 'love of oranges [gen]').

2. If the predicate noun belongs to the class of nouns of object (i.e. such nouns as *sovet* 'council', *predloženie* 'suggestion', *namerenie* 'intention', *vospominanie* 'recollection'), the subordinate genitive of such a noun is interpreted, non-ambiguously, as referring to the subject:

(7) predloženie *Kasparova* 'suggestion of Kasparov'

This is but natural – in fact, disappearance of the participant object is inherent in the essence of their derivation pattern: Object valence of the noun is filled, so to say, by the noun itself.

3. If the name has a REDUCED DIATHESIS (in the sense of Padučeva 1977), namely, objectless, the dependent genitive is unambiguously interpreted as the genitive of subject, the implied object being expressed somewhere in the context (in example below the implied object is underlined):

(8) Mašina ne prošla proverki *èkspertov* 'the machine didn't undergo the check of *experts*'
On poddalsja na obman *zlopyxatelej* 'he yielded to deception of *scoundrels*'
Èto obespečilo emu podderžku *kolleg* 'it ensured to him support *of the colleagues*'

On the other hand, if the noun has a genitive expressing the object then the genitive subject is excluded in principle (see Padučeva 1984) – the only case possible for the subject is the instrumental, as in (4).

As for possessives, they can express subject in the context of a predicative noun with the genitive object, see impossible (3c) and flawless (3a) Example elucidating this distinction between genitive and the possessive pronoun was given in (Iordanskaja 1967: 23).

It is remarkable that all possessives behave in the same way – not only possessive pronoun is acceptable in the context of a genitive subject but also the possessive adjective:

(9) gedelevo *dokazatel'stvo* teoremy o polnote 'Goedel's *proof* of this theorem'
papino *istolkovanie* moej pros'by 'father's *interpretation* of my request'
šaljapinskoe *ispolnenie* ètoj arii 'Shaljapin's *performance* of this aria'

Thus, examples (3) and (9) exemplify what may be called a POSSESSIVE-GENITIVE DIATHESIS of a deverbal noun: possessive-subject, genitive-object. The problem is that this diathesis isn't possible for all predicate names. For example, (10a), (11a) are not correct – instead one should say (10b), (11b):

(10) a. **moe* sobljudenie tajny '*my* keeping of the secret'
b. sobljudenie *mnoju* tajny 'keeping of the secret *by me*'

(11) a. *Vse zavisit ot *ego* sobljudenija tajny 'everything depends on *his* keeping of the secret'
b. Vse zavisit ot sobljudenija *im* tajny 'everything depends on the keeping of the secret *by him*'

The class of nouns compatible with the possessive-genitive diathesis is to be explored. In fact, in the context of this class the construction in question is not only possible – in fact, possessive is here the only possibility to express the subject.

In (Padučeva 1984) names affording the diathesis with the possessive subject and genitive object were attested as MODE OF ACTION nouns. Such nouns are foreseen in the classification of predicative nouns in "Meaning – Text" model. But examples are given, both by Mel'chuk and by Apresjan, only of nouns either derived from intransitive verbs (*poxodka* 'step', *povedenie* 'behavior') or having a generic object denoting a habitual action or even a property ensuing from habitual actions (*počerk* 'handwriting', *proiznošenie* 'pronunciation'). For such nouns possessive-genitive diathesis is, naturally, impossible.

At the same time, in (Apresjan 1974: 199), where regular ambiguity is discussed of nouns of action and nouns of manner of action, such nouns as *perevod* 'translation', *redakcija* 'edition' are mentioned as mode of action nouns. It is also noted that syncretism is possible of mode of action nouns and nouns of result. If so then I don't bother if some of the nouns disposing of possessive-genitive diathesis, can be called both manner of action nouns and nouns of result: a strict border is here absent.

A class of mode of action nouns outlined in this way is sufficiently broad. Mode of action nouns motivated by transitive verbs are of special interest (such as *traktovka*, *ispolnenie*, *ponimanie*, *osmyslenie*, *izobraženie*, *upotreblenie*, *postanovka*, *tolkovanie*). In fact, these are nouns having subject and object in their argument structure (Grimshaw 1990) (i.e. "glubinno-syntaksičeskaja" structure), though they are not situation names, which, according to the theory, preserve the main valences of a transitive verb.

Mode of action nouns are formed from verbs with the valences "Who?", "Whom/What?" and "How?"; it is this last valence that differentiates mode of action nouns from situation nouns.

The verb-noun derivation pattern looks like this:

X *traktuet* Y Z-*ovo* ⇒ X-*ova traktovka* Y-*a* Z-*ova*.

'X treats Y in a Z way ⇒ X's treatment of Y is Z-like'.

As is known, the general rule says that the deverbal noun denoting the *i*-th argument of the verb V fills its *i*-th valence by itself, so that this valence is cancelled in the argument structure of the name. Are mode of action nouns an exception to this rule? In fact, they seem not to loose any valences.

The explanation is to be looked for in the paradoxical status of the argument Result. In fact, in many verb classes the position of direct object is occupied by the participant Patient (for example: *prokolol podošvu* 'pierced the sole') or some other participant, such as Goal, as in *rešil zadaču* 'solved the problem' or Source, as in *otrecenziroval stat'ju* 'reviewed an article'. While the participant Result does not enter the argument structure of the verb. When the name of result is formed the participant Result can be said to be excluded from the argument structure of the noun, but it has no effect on the syntactic potential of the word.

Hence this remarkable fact that mode of action nouns (being at the same time nouns of Result, preserve the main valences of the motivating verb, both subject and object one).

So, mode of action nouns inherit from the transitive verb both subject and object valences because they are formed along the derivation pattern that doesn't absorb either subject or object. Still they differ from situation nouns, for their subject valence cannot be saturated by an NP in the instrumental but only by a possessive pronoun or a possessive adjective (such as *mamin* 'mother's', *Andreev* 'Andrej's') – including the adjective in *-ovskij* (*otcovskij* 'father's', *Rasselovskij* 'Russel's'). Some examples (from Padučeva 1984).

(12) Ego *vybor* sekundanta byl neudačen 'his choice of a second was unsuccessful'
Č'e *ispolnenie* Šopena Vam ponravilos' bol'še? 'whose rendition of Shopin you liked most'
A kakoe Vaše *ob`jasnenie* pričin ètoj ssory 'and what is your explanation of the reasons of this quarrel?'

(13) *rasselovskaja* traktovka sobstvennyx imen 'Russel's treatment of proper names'
Vot kak obstoit delo v *maminom* ponimanii 'this is how the matter stands in mother's opinion'

If the subject valence is filled by instrumental the noun cannot be interpreted as a mode of action noun:

(14) а. Ego *ponimanie* ètoj problemy ne sovpadaet s moim 'his *understanding* of the problem doesn't coincide with mine' [*понимание* 'understanding' – mode of action name]
б. *Ponimanie* im ètoj problemy svidetel'stvuet o ego iskušennosti v takix delax 'the fact that he understands the problem testifies his experience in the matters like this one' [*ponimanie* – name of a fact]

(15) а. Ego *ispolnenie* Šopena bylo velikolepno ‘his *performance* of Shopen was magnificent’ [*ispolnenie* – mode of action name]
б. *Ispolnenie* im Šopena bylo neumestno ‘*performance* of Shopen by him was out of place’ [*ispolnenie* – name of action]

A mode of action noun can be a noun of argument and a noun of circumstance. Such words as *počerk*, *poxodka* have the idea ‘How’ in their semantics, and the corresponding participant is inherited from the circumstance of the motivating verb. The noun *udar* ‘stroke’ in the context *U nego sil’nyj udar* ‘he has a strong stroke’ (example from Mel’čuk 1974: 87) is also the name of a circumstance. NB also such nouns as *povedenie* ‘behavior’, *proiznošenie* ‘pronounciation’.

It is very often the case that the participant subject is not inherited by a circumstantial noun, cf. such deverbal nouns as *vyxod*, *vyezd*. Nevertheless it is not a general rule inevitable in the course of semantic derivation.

The importance of making the difference between situation and mode of action nouns can be demonstrated on the following example:

(16) *Opisanie Tomsonom uslovij, opredeljajuščix vybor padeža, ostaetsja odnim iz lučšix ‘description by Thomson of conditions determining the choice of case is up till now one of the best’

The NP *opisanie Tomsonom uslovij, opredeljajuščix vybor padeža* is well formed and causes no doubts. But it is out of place in this context – in fact, the noun *opisanie* is assigned a diathesis appropriate for nouns of event or fact, though in the context *ostaetsja odnim iz lučšix* it should be interpreted as a mode of action noun, so that the subject is to be a possessive and not instrumental:

(17) Tomsonovskoe opisanie uslovij, opredeljajuščix vybor padeža, ostaetsja odnim iz lučšix ‘Thomson’s description of conditions determining the choice of case is up till now one of the best’

The opposite is also true – if a noun cannot be understood as a mode of action noun a possessive cannot be used as its subject marker. In (18) for example, a possessive pronoun cannot be substituted for the instrumental because the name *ponimanie*, in this context, is not a mode of action noun:

(18) Neobxodimoe uslovie perevoda – ponimanie eju <mašinoj> teksta v polnom ob`eme ‘a necessary condition of adequate translation is a complete understanding by it <the machine> of the text’

No wonder that mode of action nouns are readily used in the context of characterizing predicate, see (17) above and (19) below:

(19) *ego izobraženie Napoleona* sliškom čelovečnoe ‘his depiction of Napoleon is too human’

Mode of action nouns have one additional specific diathesis with the genitive object raising – this diathesis can be called attributive:

(20) а. ego izobraženie *Napoleona* ‘his depiction *of Napoleon*’
б. *Napoleon* v ego izobraženii ‘*Napoleon* in his depiction’

Note that the meaning of possessive adjectives in *-ovskij* and *-in* is not identical: the NP *papina fotografija* ‘father’s photo’ has two interpretations (*papa* ‘father’ the subject and *fotografija* ‘photograph’ a mode of action noun, and *papa* the object), while *velaskesovskij portret* ‘Velasquez’s portrait’ has only one meaning (Velasquez – the subject).

5 Examples

Examples below are given in order to demonstrate two diatheses of a mode of action noun – possessive-genitive and attributive. Examples are given for words *analiz* 'analysis', *videnie* '*perception*', *vybor* 'choice', *izobraženie* 'depiction', *izloženie* 'account', *koncepcija* 'conception', *ocenka* 'evaluation', *perevod* 'translation', *plan* 'plan', *podbor* 'selection', *ponimanie* 'understanding', *postanovka* 'staging', *pročtenie* 'reading', *rešenie* 'solution', *sposob* 'mode', *traktovka* 'treatment', *tolkovanie* 'explication', *formulirovka* 'formulation'.

ANALIZ
On otmečaet, čto *saxarovskij analiz opasnosti* mirovogo jadernogo samoubijstva v točnosti sovpadaet s dokumentami, opublikovannymi amerikanskimi učenymi.

VIDENIE
… perenos v sovremennost' dejstvija romana Oruèlla liš' banaliziruet tragediju *oruèllovskogo videnija totalitarizma* XX veka...

VOSPRIJATIE
Èti otnošenija pomogajut vyjavit' nekotorye črezvyčajno suščestvennye aspekty *šoloxovskogo vosprijatija žizni*.

VYBOR
(1) Vidimo, kontakt s aviapromyšlennost'ju i poval'naja uvlečennost' aviaciej 30 godov skazalis' na *otcovskom vybore* buduščej *professii* syna.
(2) rabotaet uže 23 goda, i ni razu ne požalela o *svoem vybore professii*!
(3) Esli čelovek lišen vozmožnosti vybora zla, to *ego vybor dobra* polnost'ju deval'viruetsja.

IZOBRAŽENIE
Naprimer, k *ego izobraženiju revnivca*, v lice Otello, nel'zja pribavit' ni odnoj čerty: tak ono polno.

IZLOŽENIE
Russkomu že čitatelju, odolevšemu *rollanovskoe izloženie* ètogo cirkovogo *vran'ja*, napomnju èto.

INTERPRETACIJA
I zdes' snova prixodit na pamjat' "Andželo" i *ego lotmanovskaja interpretacija*.
Besy, ved'my i privoroty – èto uže sugubo *otcovskaja interpretacija.* [Anna Tkačeva. Privorot (1996)]

ISPOLNENIE
proslušivalas' magnitofonnaja zapis' "Petuškov" v *Veničkinom ispolnenii*
No, požaluj, imenno *Runino ispolnenie* do six por predstavljaetsja mne naibolee točnym.

ISTOLKOVANIE
čuvstvo jumora ej nevedomo – čto ona dokazala *svoim istolkovaniem slova* "trubadur"

KONCEPCIJA
To, čto s nim tam proisxodit, vpolne ukladyvaetsja v *bessonovskuju koncepciju* francuzskogo kino.
približalsja k *tolstovskoj koncepcii iskusstva*
No vernemsja k *nemcovskoj koncepcii pravitel'stva-uborščicy*.
platonovskaja koncepcija ljubvi kak vosxoždenija po lestnice prekrasnogo
V *Darvinovskoj koncepcii* èvoljucija sostoit v tom, čto imeet mesto mexanizm prisposoblenija.

OCENKA
A nas vsex v klasse porazila *ego ocenka* moix skromnyx *uspexov*.

PEREVOD
V *pasternakovskom perevode Šekspira*, kak i v tekste Griboedova, postojanno proisxodit peremol vtorgajuščixsja v nego kuskov prozy.
Ljuterovskij perevod "Novogo Zaveta" (1522) markiruet načalo principial'no novogo otnošenija k Pisaniju.

PLAN
čubajsovskij plan restrukturizacii RAO "EÈS Rossii" prokremlevskie frakcii poručili provodit' ne "komande" samogo Čubajsa, a pravitel'stvu M. Kas'janova
V Pariže car' zakazal konduktoru Djumenju kopiju *leblonovskogo plana Peterburga*, zaplativ za nee 100 livrov.
Ustinovskie plany ukreplenija peredovoj fenomenal'ny po masštabam, po raznoobraziju primenjaemyx sredstv i detal'nosti prorabotki vsego ètogo raznoobrazija.

PODBOR
Ja lovlju sebja sejčas na tom, čto *moj podbor svidetel'stv* i argumentov sliškom tendencioznyj, odnostoronnij, umyšlennyj.

PONIMANIE
Nravstvennye osnovy byli v menja založeny v sem'e, tak čto *moe ponimanie grexa* – ne cerkovnoe, a žitejskoe.
ja … starajus' dat' sebe otčet, čto novogo vnesla èta fraza v *moe ponimanie teksta* i kak perestroila staroe
Esli èto *moe širokoe ponimanie slova* revoljucija pokazalos' g. Astaf'evu nepravil'nym, to on mog by prjamo na èto vozrazit'...
… oppozicionnuju kul'turu, po svoemu smyslu blizkuju k kul'ture kritičeskogo *diskursa* v *gouldnerovskom ponimanii*
jungovskoe ponimanie duševnoj *žizni* čeloveka "kak vnutrennej dramy so množestvom personažej"

POSTANOVKA
Mandel'štam, očevidno, ulovil karnaval'nuju storonu revoljucionnogo "dejstva", načavšegosja s *mejerxol'dovskoj postanovki "Maskarada"*.
Čisto *mandel'štamovskaja postanovka problemy* povedenija.

PROČTENIE
Stol' že somnitel'no *ego pročtenie* privatizacii.
S *"nabokovskim" pročteniem Gogolja* možno ne soglašat'sja.

REŠENIE
Svoe rešenie problemy predlagajut neskol'ko amerikanskix kompanij, kotorye vypuskajut kamufljažnye kontejnery dlja bazovyx stancij.
najti edinstvennyj ključ k ponimaniju rasskaza ili daže najti *čexovskoe rešenie voprosov*, postavlennyx v nem

SPOSOB
èto – liš' *lužkovskij sposob ogradit'* rossijan ot raspredelenija obščenacional'nogo doxoda

TRAKTOVKA
Kommunističeskie bjurokraty vozmuščalis' *mejerxol'dovskoj traktovkoj "Revizora"* – nezamaskirovannoj satiroj na nix samix.

TOLKOVANIE
Interesno, … čto oni dumajut o *moem tolkovanii Apokalipsisa*?

FORMULIROVKA
Ibo ètoj "svobody duxa" nikogda i ne moglo byt' u nas imenno iz-za pervyx dvux položenij *aksakovskoj formulirovki "russkogo puti"*.

6 Conclusion

Thus, a class of predicative nouns is presented (namely, mode of action nouns) that have definite common character and a specific diathesis with subject-possessive. The derivation model that generates possessive adjectives in *-ovskij* is highly productive and neologisms of this kind are widely used in modern Russian.

I am not sure that the term mode of action nouns is appropriate for the class of nouns in question but the class described definitely deserves attention.

References

Apresjan 1974: Апресян Ю. Д. 1974. *Лексическая семантика: Синонимические средства языка*. М.: Наука.

Grimshaw J. 1990. *Argument Structure*. L. etc.: MIT Press.

Iordanskaja 1967: Иорданская Л. Н. 1967. *Автоматический синтаксический анализ*. Наука, Сибирское отделение. Новосибирск.

Mel'čuk 1974: Мельчук И. А. 1974. *Опыт теории лингвистических моделей «Смысл ↔ Текст». Ч. 1. Семантика, синтаксис*. М.: Наука.

Padučeva 1974: Падучева Е. В. 1974. *О семантике синтаксиса*. М.: Наука.

Padučeva 1977: Падучева Е. В. 1977. О производных диатезах отпредикатных имен в русском языке. *Проблемы лингвистической типологии и структуры языка*. Л.: Наука: 84–107.

Padučeva 1984: Падучева Е. В. 1984. Притяжательное местоимение и проблема залога отглагольного имени. *Проблемы структурной лингвистики*. М.: Наука: 50–66.

Shmelev 2008: Шмелев А. Д. 2008. Посессивы в современной русской грамматике. *Динамические модели. Слово, предложение, текст*. М.: Языки славянских культур: 927–942.

Zemskaja 1992: Земская Е. А. 1992. *Словообразование как деятельность*. М.: Наука.

Lexical units and syntactic constructions: the caused-motion construction

Marta Rebolledo Lemus
Universidade da Coruña
marta.rebolledo.lemus@gmail.com

Margarita Alonso Ramos
Universidade da Coruña
lxalonso@udc.es

Abstract

This work aims at discerning to what extent syntactic differences in the use of a word must lead to the distinction of different lexical units. To study this question, we examine the relation between the caused-motion construction and lexical units that can enter into in two theoretical models: FrameNet and Meaning Text Theory.

1 Introduction

This paper focuses on the relation between lexical units and the syntactic constructions they can accept. The objective is to evaluate if the different syntactic constructions a word can bear are a strong enough criterion to distinguish lexical units. This problem is especially relevant when dealing with the different senses of a verb. The study of verbal alternation developed by Levin (1993) is a good example of the importance of the contexts in the delimitation of verbal classes. From a different point of view, Van Valin & LaPolla (1997), among others, maintain that the semantic properties of predicates, especially verbs, establish the core semantic representations of a clause or a sentence. From this perspective, the problem of how the arguments of predicates are linked to syntax is one of the central questions (Tenny 1994; Van Voorst 1988, among others). The Construction Grammar approach (Fillmore et al. 1988, Goldberg 1995, 1996, Jackendoff 1997) as well as other approaches, offers a different solution to this problem. Given that its basic proposal is that syntactic constructions bear independent meanings, it transfers the problem from the delimitation of different verb senses to the description of the relation between constructions and lexical units.

In this work we concentrate on how the relation between syntactic context and lexical units is treated in two theoretical models: FrameNet (FN) and the Meaning Text Theory (MTT). More specifically, we examine the analysis these two models offer, to describe the so-called *caused-motion construction* (Goldberg 1995), as it can be seen in the following examples:

(1) a. *Bob sneezed*
b. *Bob sneezed the napkin off the table*

(2) a. *Bob laughed*
b. *Bob laughed his son out of the room*

In this alternation, a generally intransitive verb, in this case, *sneeze* or *laugh*, adopts a transitive realization with a locative element. The use of this syntactic construction is associated to another meaning that joins the meaning of the verb and the meaning of movement: roughly, 'sneeze / laugh of X causes Y moves from/to Z'. One theoretical possibility could be to establish two different lexical units, a transitive and an intransitive one, so that they both have their own lexicographic entry. Another possibility would be assigning the change of meaning to the syntactic construction, as it is suggested by Construction Grammar (Goldberg 1995). As we will see, none of these explanations seem totally adequate. In what follows, we will show the solutions given by FrameNet and the TST.

There are important differences between the two models revised in this work. One of the most important for our concerns is that establishing word senses is not one of the objectives of FrameNet. However, this project is engaged in giving an account of all the syntactic information of a lexical unit. Therefore, we think that the study of the treatment of the causative-motion construction in both models will lead us to evaluate the terms in which the relation between lexical units and syntactic constructions is posed.

2 The Caused-motion construction in FrameNet

As it is known, the original objective of the FrameNet project is the description of lexical units in terms of the semantic frames they evoke, as well as the description of these frames themselves. Besides, the FrameNet research aims at defining the range of combinatorial possibilities as valence patterns. This implies the description of the syntactic realization of each lexical unit:

> The goal of FrameNet lexical descriptions is, for each frame-bearing word, to match the word's semantic combinatorial requirements with the manner of their syntactic realization (Fillmore 2008, 51-52).

As Fillmore et al. (2002) point out, an efficient disambiguation depends on the information about the combinatory of each word for each of its senses, so it becomes fundamental that the models reflect this combinatory as precisely as possible.

In FrameNet, predicates belong to frames based on their shared semantics, not on similar syntactic alternations as in Levin (1993) (Baker & Rupenhoffer 2002). This means that FrameNet may be able to reflect the fact that verbs with similar semantics show different syntactic patterns, but as we will see, it has difficulties in presenting the peculiarities of verbs with a particular behaviour.

The Lexicon contained in the initial project is being complemented by the elaboration of a *constructicon* defined by Fillmore (2008) as 'a record of English grammatical *constructions*'. This project is aimed at the labelling of constructions which ordinary parsers are not likely to notice or which grammar checkers are likely to question:

> Some of them involve purely grammatical patterns with no reference to any lexical items that participate in them, some involve descriptions of enhanced demands that certain lexical units make on their surroundings, and some are mixtures of the two (Fillmore 2008: 49).

But the analysis of these units shows a wide range of problems. Fillmore (2008) displays 21 constructions which present annotation difficulties. Some of them are the following (using his labels):

1. Lexical constructions*: sneeze the napkin off the table*
2. Verbs with contextual requirements outside of their phrasal projection: *it's too dark to tell what they're doing*
3. Templatic constructions: *Six is to three as four is to two*
4. Presentative constructions: *Here she is*
5. Constructions similar to *Wherewithall*: *I don't have the resources to landscape the garden*, [John]…*who will provide me the wherewithal to accomplish this* (…).
6. Gapping: *John loves, but Mary hates, rock music*
7. *Let alone*
8. Verb one's way: *Let's start making our way home*.
9. In one's own right: *The son of a poet can be a fine poet in his own right*.
10. Measurement phrases*: five meter long/wide*.
11. The + Adjective*: the rich, the poor*.
12. Adjective comparison*: she's much more intelligent than you said.*
13. (...)

The variety of units shown in this list is derived from the fuzziness of the notion of *construction* itself. Following authors such as Fillmore & Kay (1999) or Jackendoff (2002), constructions are defined with a high degree of abstraction:

> A construction (e.g. the subject-auxiliary inversion construction) is a set of conditions licensing a class of actual constructs of a language (Fillmore & Kay 1999, 3).

Langacker (1987, 25-27) identifies every grammatical unit, from morphemes to syntactic structures, as constructions. He proposes a lexicon-syntax continuum based on this notion. Fillmore (2008) assumes both the definition and the continuum: 'I count myself among the linguists who believe in a continuity between grammar and lexicon' (Fillmore 2008, 49).

The construction-building work includes various kinds of idioms or other multiwords. Fillmore (2008) points out that FrameNet tools are enough when describing phrasal verbs (*pick up, take up*) or words with selected prepositional complements (*depend on, fond of, interest in*). However, with the same tools, they find difficulties in describing and annotating the constructions in the corpus; in contrast to the lexicographic annotation that is linked to a target lexical item, in the constructional annotation there is no target lexical item to link the construction to (Fillmore 2008, 59).

As for the caused-motion construction, the relation with the lexical unit is not clear, as the change in the verbal valence and its associated meaning can be attributed to the construction itself. More explicitly, the use of *sneeze* in this construction implies the use of two arguments that lack a corresponding frame element:

(3) *Bob sneezed the napkin off the table*

Usually, this verb evokes a frame where there are no elements such as Theme or Goal. The only way to describe this pattern and its meaning would be relating this *sneeze* to another frame as Cause-Motion, evoked by verbs such as *push* or *throw*. In this frame we can find the Theme and the elements related to the described movement (as Goal, Area, or Source) as core elements. For instance,

(4) *She would not throw her coin into the Trevi Fountain*

It might be questionable whether we really face two different phenomena in *sneeze* (3) and in *throw* (4). It can be said that in examples such as (3), the element *off the table* has the role Source, in a similar way to *into the Trevi Fountain* in the previous example represents the Goal of *throw*. However, to say that this *sneeze* (3) evokes another frame entails the creation of another lexical unit for this meaning, which goes against the constructional approach that avoids the multiplication of lexical entries. Fillmore (2008, 60) leaves this question open when he points out that there is no automatic way of deciding whether the lexicographer must list in the lexicon the behaviour of the word used in the construction (*sneeze*, in this case) or he must merely recognize it as an instance of the construction.

This question does not arise for other practitioners of the constructional approach, such as Goldberg (1995). She argues against the derivation of constructional meaning from the individual elements of the construction (the verb and the preposition), and thus she asserts that a construction must be posited in the grammar. For this author, the verb carries the specific meaning, and the construction adds another part of the meaning ('X causes Y to move Z'). In this case, the Theme and the Goal of *sneeze* are provided by the construction, and thus there is no need to create a new lexical entry. Besides, a sort of compatibility between the semantics of the verb and the construction is necessary; as Goldberg (1995, 166) shows, the construction imposes certain semantic restrictions on the verbs that can enter into it:

(5) a. *Pat coaxed him into the room.*
b. **Pat encouraged/convinced/ persuaded/instructed him into the room.*

What all of the verbs in (5b) have in common is that they entail that the entity denoted by the direct object makes a cognitive decision. However, the caused-motion construction imposes as a semantic constraint that no cognitive decision can mediate between the causing event and the entailed motion.

Goldberg points out other constraints that serve to characterize semantically the construction. Nevertheless, she does not give any indication about how the caused-motion construction could be registered in a constructicon.

It seems that the difficulties shown by FrameNet when describing the caused-motion construction can be explained by four related aspects of the approach examined:

a. The conceptual base of the frames. Even though it allows for stating the semantic similarities of a group of verbs, it disallows to reflect all the different syntactic patterns a verb can bear.
b. The lack of precision in determining all the frames a verb is related to.
c. The fuzziness of the notion of *construction*, which leads to the following point:
d. The indetermination of the structure of the possible constructional entries. More precisely, it should be indicated how they will include the relevant information about the lexical items that participate in them.

3 The Caused-motion construction in the ECD

The lexicon-syntax continuum lying under the model illustrated by Fillmore (2008) doesn't hold in the Meaning Text Theory (MTT) and the *Explanatory and Combinatorial Dictionary* (ECD). As it is known, for this model the central aspect of language is the lexicon. This centrality, far from avoiding the differentiation of linguistic domains, strengthens their distinction: 'a lexicon of L describes L's individual lexical signs, and the grammar of L covers a) L's individual grammatical signs, and b) the behaviour of sets of L's signs' (Mel'čuk 2006, 3).

This will to establish boundaries, which is rejected by Construction Grammar, allows to describe in a dictionary many of the entities Fillmore cannot give account for (especially, multiword units of the kind of *let alone*). However, the main advantage is not the possibility of including more multiword units in the lexicon, but the highest precision the ECD seems to achieve in the identification of the semantic variations of a word in relation to its morphological properties and its syntactic realizations. This highest precision is acquired by a set of rules used to describe in which cases and in what manner different contexts modify word meanings. Basically, in the ECD, formal clues of semantic differences are observed in three well-defined domains (Mel'čuk 2006, 295):

1) Morphological properties (e.g., different inflection patterns for different L' uses)
2) Government Pattern (different means for the expression of actants with different L' uses)
3) Semantic derivations and collocates (paradigmatic and syntagmatic lexical relations)

The analysis of these three levels related to a LU allows for a decision as to whether we have to split a lexical unit (LU) into two or maintain just one. With this purpose, we will evaluate the weight of the formal differences found. In other words:

> If the semantic difference between two uses of L is correlated with two subsets I1 and I2 of differentiating lexicographic information which show more than one formal difference, then L should be split in two LUs L1 and L2 (Mel'čuk 2006, 295).

That is to say, within the most rigid functional tradition, different lexical units are distinguished only if there are enough formal differences. On the contrary, the absence of enough formal differences will indicate that the semantic differences being considered are not strong enough to split a lexical unit:[1]

This methodological procedure stands out due to its rigorousness and descriptive precision, but it is far from simplifying the huge question. In fact, it illustrates the thin gradation between a contextual difference and a change of lexical unit, by signaling that a different constructional pattern not always leads to a different lexical unit: the same lexical unit can present different constructional patterns, and we must describe these possibilities. To endeavour this task, the MTT has developed an exhaustive

[1] Some examples of formal differences and their incidence in the delimitation of lexical units can be seen in Mel'čuk (2006 76-77). For example "the case of AUNT: taken in any of its three possible uses, this noun has the same morphology, syntax and cooccurrence", and this allows to conclude that we are in front of only one lexical unit.

theoretic mechanism that will be applied here to the cases of the caused-motion alternation, already seen in examples (1) and (2) and now exemplified with the verb *dance* (6):

(6) a. *Bob danced.*
b. *Bob danced her into the corridor.*

As we have already seen in the previous section, verbs admitting this construction undergo a variation on their actantial structure. While in the (a) examples we face a verb with only one actant, referring to a situation with only one participant, the (b) examples refer to a situation[2] with three participants: the agent causing motion, the object moved, and the goal or the source. The question is to what extent this alternation implies the creation of a new LU.

In order to provide an answer to this question, it must be noted that this construction, although not completely predictable, is quite regular in English. As Mel'čuk (forthcoming) has pointed out, "not all theoretically possible conversional derivations of this type are actually acceptable to speakers". As we have said before, whereas you can *coax someone onto somewhere*, it is not possible **to encourage someone onto somewhere*. Apparently, this unsystematic behaviour could be attributed to the typical properties of derivation processes. As these transitive verbs are considered to be derived LUs that share with the basic LU (the intransitive one) only the first actant, they could have their own lexical entry. But apart from implying an unnecessary duplication of lexical entries, this solution would not reflect the partially regular character of this derivation process.

In fact, LUs which are regularly derived do not need a lexical entry, because they are not considered *actual* lexical units, but *potential* lexical units. In MTT a potential unit is a lexical unit that can be created in an almost systematic way from an actual lexical unit, which is a unit with a lexical entry. These potential LUs are divided into two major classes: compound potential LUs and derived potential LUs.

A compound LU is built out of two or more actual LUs (as in *Chinese-born*). A derived lexical unit can be affixally derived or conversionally derived. An affixally derived lexical unit is built out of an actual LU by a derivational affix, as in *amorphous+ness*. A conversionally derived lexical unit is built out of an actual LU by changing its syntactic valence, so that the LU is suited to be used in a particular syntactic construction, in which it cannot be employed without the aforementioned derivation, as in *Bob sneezed the napkin off the table* or *Bob was rumored into this marriage*. Therefore, the transitive *sneeze* is a derived LU regularly created through a derivational semantic rule. The rule, which describes the caused-motion sense, will have the following general form (Mel'čuk 2007):

(7)
'by affecting Y, action P of X
causes1 Y to move
from Z to W over the trajectory T' <==> UL**n^{o}** ('P') | with modification of the GP

In the particular case of *sneeze*, it is stated as follows:[3]

(8)
'by affecting Y, sneeze of X
causes1 Y to move from Z ' <==> SNEEZEn^{o*}[4]| Add two columns to the GP of sneeze 1

This rule describes the meaning of a derivational means, that is, a derivateme. The derivateme under consideration is expressed by a conversion that changes the government pattern (GP) of the basic lexical unit of the vocable (Mel'čuk forthcoming). This rule reflects the possibility of creating a lexical unit with three arguments from the one-argument *sneeze*. As the potential LUs are not listed as

[2] About the notion of linguistic situation, see Mel'čuk 2004 and Alonso Ramos 2007.

[3] The [cause1] value denotes no agentive causation, opposite from [cause2].

[4] This asterisk indicates the provisional or ephemeral character of any potential LU.

such in the dictionary, the GP of the derivate LU is described as a condition of the derivational semantic rule. For example, the potential LU SNEEZE [N_Y *off* N_Z], adds the following two columns to the lexicographic GP of its underlying lexical unit:[5]

(9) GP added to SNEEZE 1

Y ⇔ **II**	Z ⇔ **III**
1. N	1. *off* N

By means of this analysis, the DEC can reflect a unit's change of syntactic structure without unnecessarily multiplying the number of entries of a word, which is one of the main problems of current dictionaries, as Battaner & Torner (2008) point out. These potential derivate lexical units are not listed in the lexicon, but are constructed by word-formation rules out of actual LUs (based, of course, on the lexical entries of the latter). We are dealing with a potential LU, specifically created through a derivational word-formation rule that allows an English speaker to use the 'normal' verb [*to*] SNEEZE in the causative motion sense.

All LUs derived by this 'caused-move' rule share the same "constructional" meaning. They distinguish themselves in the instantiation of the predicate P ('sneeze', 'laugh', etc.) and in the movement arguments chosen: some choose the path, as *dance* (*her across the corridor*), some choose the goal, as *dance* (*her into the corridor*). However, the rule does not give account of the difference between *coax* and *encourage*, that Goldberg (1995) has pointed out. Let's recall the examples:

(10) a. *Pat coaxed him into the room.*
b. **Pat encouraged/convinced/ persuaded/instructed him into the room.*

Goldberg (1995) chooses this example, among others, to argue against the claimed idiosyncrasy that would force the multiplication of lexical entries. As we have already mentioned, she explains that there is no idiosyncrasy but a semantic reason: "no cognitive decision can mediate between the causing event and the entailed motion" (Goldberg 1995, 167). The MTT derivational semantic rule would be able to describe this difference by the addition of a semantic component, as we propose here (marked in bold):

(11)
'by affecting Y, action P of X
causes1 Y to move
from Z to W over the trajectory T
without Y makes a cognitive decision' <==> UL**n°** ('P') | with modification of the GP

With this added semantic component, the rule prevents the possibility of deriving *encourage* or *convince* with the meaning of cause-motion without the necessity of lexical stipulations, as criticised by the constructional approach. However, even with this formulation the cause-move derivation rule does not cover all the possibilities. Thus, it does not give account of the derivation from transitive verbs, such as *kick the dog into the bathroom* or *break the eggs into the bowl*. The rule should be generalized to cover these cases, or should be split in several more specific rules. This decision will depend on a deeper study, but what counts here is that in MTT the means to describe thoroughly this construction exist.

4 Resemblances and divergences between the two approaches

In the previous sections, the treatment of the caused-motion construction in both approaches has been presented. It has been shown that FrameNet presents certain difficulties in treating adequately the fact that verbs such as *sneeze* or *laugh* enter in two constructions without creating a new lexical entry. It is

[5] To maintain the caused-motion sense, the preposition introducing the expression of DSyntA **III** is semantically full, and it has to be present in the DSyntS.

still not clear how the Construction Entries will look, where they will include information about lexical items that participate in them, and more particularly, how the caused-motion construction entry will look like. However, it must be highlighted that Fillmore's linguistic conception in this respect is not very different from Mel'čuk's: both assume a dynamic creation of lexical units such as *sneeze* transitive, instead of creating a new lexical entry. For MTT, as we have just seen, the depicted derivational semantic rule describes the meaning associated with the syntactic valence change without opening a new lexical entry for SNEEZE. In a similar way, Fillmore (2008, 67) explicitly mentions the generation of lexical units:

> Some products of a construction are simply lexical units in essentially every way, except in that they are "generated" rather than requiring individual listing in a dictionary's wordlist: this is true of the products of argument structure constructions.

Another resemblance between both approaches is the importance given to the lexicon. Both Fillmore and Mel'čuk refuse the old vision of the lexicon as a grammar appendix, and it is not fortuitous that both have undertook the task to compile a lexicon or dictionary. Mel'čuk could also sign the following words, written by Fillmore (2008, 49): "each lexical item carries with it instructions on how it fits into a larger semantic-syntactic structure, or, alternatively, on how semantic-syntactic structures are to be built around it".

Differences appear in the relation between lexicon and grammar. As we said before, Fillmore speaks in terms of "continuity between grammar and lexicon", whereas, for Mel'čuk, the lexicon primes logically on the grammar. This different conception of grammar may be the source of the diverse treatments of linguistic phenomena observed in the two approaches. The constructional approach that nourishes the constructicon proposes the vague notion of *construction* for very different phenomena, as if the notion of construction in itself was explanatory. On the contrary, MTT offers different solutions to the so-called constructions.

In this paper, we have focused on the treatment of the caused-motion construction by a derivational semantic rule. However, not all of the argument structure constructions would be treated by derivational semantic rules. For example, the description of the so-called *ditransitive construction*, as in *Mary made John a cake* strongly differs from the one of caused-motion constructions. In this case, we do not have a potential LU *make*, with a "special meaning", but a verb that allows the realization in surface syntax of an Indirect Object not predicted by its government pattern. The Beneficiary Indirect Object (*John*) corresponds not to a deep syntax actant, but to a circumstantial. At this moment, Mel'čuk (forthcoming) proposes the uses of deep fictitious lexical units to represent at the deep syntax level the fact that a lexical-type meaning is expressed by a syntactic construction (i.e. a surface-syntactic relation or a configuration of such relations). In the case of Beneficiary Indirect Object, the fictitious lexical unit is the preposition "FOR".

In FrameNet, the description of this Beneficiary Indirect Object is simple. Verbs such as *bake* or *make* evoke the Cooking-Creation Frame, where the Recipient is an Extrathematic Frame Element. For example, in the entry for MAKE, we find the following examples with Recipient:

(12) a. *Phoebe MADE all three of them hot chocolate.*
b. *We'll MAKE a big pot of tea for all of us.*

The Recipient can be realized as prepositional phrase with "for" or as an object, without more distinction. In contrast, in MTT, the deep syntactic representation of (12a) and (12b) will be different: for the first one, a fictitious "for" is needed, whereas for the second one, we have the genuine semantically full preposition "for". Evidently, they will be different in surface syntax: for the first one, we have an Indirect Object, but not for the second one, where we have an oblique object.

If the caused-motion construction is treated in MTT by a semantic derivational rule, and the ditransitive construction is treated by a fictitious lexical unit, there are transitive-intransitive alternations, such as in the transitive/intransitive verb *to rise,* for which MTT does not propose a special solution, but simply the splitting of lexical entries. While only one lexical entry is established for *sneeze*, (being the cause-motion sense obtained by a derivational rule), two different lexical entries are suggested to explain the behaviour of the verb in examples such as the following:

(13) a. *Indonesian government rises electricity price this year.*
b. *Electricity prices rise.*

The alternation exemplified in (13) is much more frequent than the *sneeze* one. But then we should wonder to what extent frequency is a parameter to differentiate lexical units (as, for instance Kilgarriff 1992 does). These kind of examples reveal the fact that the creation of a new lexical entry implies a wide range of different criteria, and it is always a gradual matter. The relevant aspect of the MTT approach is that it allows for evaluating if a new LU must be considered in relation to the formal differences observed.

5 Conclusion

The analysis based on the theoretical premises of the ECD shows that not every syntactic difference leads to the creation of a new lexical entry. The distinction between actual and potential lexical units, together with the system of derivational rules, makes it possible to describe the actantial modification observed when *sneeze* is constructed in a caused-motion sense, without the necessity of creating a new lexical entry.

In FrameNet, the grounding of the frame elements on the conceptual frames allows to reflect the semantic similarity of a group of verbs with dissimilar syntactic patterns (opposite from Levin 1993), but disallows to reflect the syntactic peculiarities of individual lexical units. The creation of a record of constructions or *constructicon* related to the lexical units of FrameNet adds the problem of the delimitation of these units. Besides, the concept of *construction* is so wide that it becomes inoperative. On the one hand, it includes a diverse variety of units and, on the other hand, it hinders from determining whether the behaviour of the elements in the constructions must be considered as part of the lexical units or as instances of the constructions they appear in.

These problems question the lexicon-semantic continuum placed on the foundations of Construction Grammar, and return us to the lexicon-syntax division as established in traditional grammar. Following this division, the MTT clearly establishes the difference between semantics, syntax and morphology. Thanks to this division, the ECD theoretical apparatus happens to be adequate to specify the level where the contextual differences of a lexical unit merge form, and consider whether they establish a new lexical unit or not. This avoids one of the main problems highlighted by Construction Grammar in 'lexical-based' theories, as it allows to establish the uses of words without unnecessarily multiplying their senses. Nevertheless, as the creation of a new LU is a gradual task, the MTT theoretical apparatus does not close the problem: questions like why do we establish two lexical entries for *rise* but only one for *sneeze* are not easily answered, and this leaves open the question of how must be reflected in the dictionary the way word meaning is modified by different contexts.

References

Alonso Ramos, Margarita. 2007. Actantes y colocaciones. *Nueva Revista de Filología Hispánica*, 55 (2): 435-458.

Baker, Collin F. & Josef Ruppenhofer. 2002. FrameNet's Frames vs. Levin's Verb Classes. In J. Larson, M. Paster (eds.), *Proceedings of the 28th Annual Meeting of the Berkeley Linguistics Society*, 27-38 [http://framenet.icsi.berkeley.edu/]

Battaner, Paz & Sergi Torner. 2008. La polisemia verbal que muestra la lexicografía. *Actas del II Congreso Internacional de Lexicografía Hispánica: el diccionario como puente entre las lenguas y culturas del mundo*. Alicante: Universidad de Alicante, 204-216.

Fillmore, Charles. 2008. Border Conflicts: FrameNet Meets Construction Grammar. In E. Bernal, J. DeCeasaris (eds.), *Proceedings of the XIII EURALEX International Congress*. Barcelona: Edicions a Petició, 49-68.

Fillmore, Charles, Paul Kay & Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language* 64: 501-38.

Fillmore, Charles J. & Paul Kay. 1999. Grammatical constructions and linguistic generalizations: The *What's X doing Y?* Construction. *Language* 75: 1-33

Fillmore, Charles J., Baker, Collin F. & Sato, Hiroaki. 2002. Seeing Arguments through Transparent Structures. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*. Las Palmas, 787-91. [http://framenet.icsi.berkeley.edu/]

Goldberg, Adele E. 1995. *Constructions. A Construction Grammar approach to Argument Structure*. Chicago: University of Chicago Press.

Goldberg, Adele E. 1996: Construction Grammar. In Keith Brown, Jim Miller (eds.), *Concise Encyclopedia of Syntactic Structures*. Oxford/New York/Tokyo: Pergamon, 68-71.

Jackendoff, Ray. 1997. Twistin' the night away. *Language* 73:534-59.

Jackendoff, Ray. 2002. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford/New York: Oxford University Press.

Kilgarriff, Adam. 1992. *Polisemy*. Brighton: University of Sussex.

Levin, Beth. 1993. *English Verbs Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.

Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar. I. Theoretical Prerequisites*. Standford (California): Standford University Press

Mel'čuk, Igor. 2006. Explanatory Combinatorial Dictionary. In G. Sica (ed.), *Open Problems in Linguistics and Lexicography*. Monza (Italy): Polimetrica Publisher, 225-355.

Mel'čuk, Igor. 2007. Semantic Transition Rules (of the Semantic Module of a Meaning-Text Linguistic Model). In K. Gerdes, T. Reuther, L. Wanner (eds.), *Proceedings of the 3rd International Conference on Meaning-Text Theory*. München, Wien: Wiener Slawistischer Almanach.

Mel'čuk, Igor. Forthcoming. *Semantics*. Amsterdam/Philadelphia: John Benjamins.

Tenny, Carol L. 1994. *Aspectual Roles and the Syntax-semantics Interface*. Dordrecht: Kluwer.

Van Valin, Robert D. & Randy J. LaPolla. 1997. *Syntax: Structure, Meaning and Function*. Cambridge: Cambridge University Press.

Van Voorst, Jan G. 1988. *Event Structure*. Amsterdam: John Benjamins.

So-Called Collective Numerals in Polish

(in Comparison with Russian)

Zygmunt Saloni

University of Varmia and Masuria in Olsztyn

Ul. Kurta Obitza 1; 10-727 Olsztyn, Poland.

saloni@uw.edu.pl

Abstract

This paper deals with Polish numeral forms of the type *pięcioro* and conditions for using them. The problem in question (as in Russian) is how to interpret them: either as forms of special ("collective") lexemes or as members of homogeneous lexemes (cardinal numerals). The author shows that these forms have basically the same meaning as cardinal numerals: they express plurality when certain nouns require them, and secondarily have the additional facultative meaning of a group of people containing both sexes. The basic function of the opposition *pięcioro*:*pięć* is similar to distinguishing genders on the basis of the subordinated word forms. The contrast with other gender forms is not very sharp (many unclear cases are shown), but the kernel of the neuter subgender defined by obligatory use of *pięcioro* forms contains one noun (***dziecko*** `child') that is very frequently used with numeral forms.

1 Introduction – Linguistic Material

In Polish, as in Russian, there exists a group of numeral forms traditionally called collective numerals. As "regular" Polish cardinal numerals, they are contrasted for case. Let us see two examples – the forms corresponding to the numbers 2 and 5:

Nom	*dwoje*	*pięcioro*
Gen	*dwojga*	*pięciorga*
Dat	*dwojgu*	*pięciorgu*
Acc	*dwoje*	*pięcioro*
Inst	*dwojgiem*	*pięciorgiem*
Loc	*dwojgu*	*pięciorgu*
Voc	*dwoje*	*pięcioro*

According to Polish traditional grammar handbooks and dictionaries, such sets of contrasted forms are treated as lexemes (for the above examples ***dwoje*** and ***pięcioro***), inflected for case, and separate from cardinal numerals (in our cases ***dwa*** and ***pięć***). In dictionaries of contemporary Polish (and – with some limitations, which we will show further – in modern usage), we can find analogous series of forms differentiated by the value of case for numbers: 2, 3 (***troje***, cf. ***trzy***), 4 (***czworo***, cf. ***cztery***), 5, 6 (***sześcioro***, cf. ***sześć***), 7 (***siedmioro***, cf. ***siedem***), 8 (***ośmioro***, cf. ***osiem***) , 9 (***dziewięcioro***, cf. ***dziewięć***), 10 (***dziesięcioro***, cf. ***dziesięć***), 11 (***jedenaścioro***, cf. ***jedenaście***), 12 (***dwanaścioro***, cf. ***dwanaście***), 13 (***trzynaścioro***, cf. ***trzynaście***), 14 (***czternaścioro***, cf. ***czternaście***), 15 (***piętnaścioro***, cf. ***piętnaście***), 16 (***szesnaścioro***, cf. ***szesnaście***), 17 (***siedemnaścioro***, cf. ***siedemnaście***), 18 (***osiemnaścioro***, cf. ***osiemnaście***), 19 (***dziewiętnaścioro***, cf. ***dziewiętnaście***), 20 (***dwadzieścioro***, cf. ***dwadzieścia***), 30 (***trzydzieścioro***, cf. ***trzydzieści***), 40 (***czterdzieścioro***, cf. ***czterdzieści***), 50 (***pięćdziesięcioro***, cf. ***pięćdziesiąt***), 60 (***sześćdziesięcioro***, cf. ***sześćdziesiąt***), 70 (***siedemdziesięcioro***, cf. ***siedemdziesiąt***), 80 (***osiemdziesięcioro***, cf. ***osiemdziesiąt***), 90 (***dziewięćdziesięcioro***, cf. ***dziewięćdziesiąt***).

Moreover, the same series exist for numerals with special meaning: ***oboje*** – for ***oba*** `both', ***obydwoje*** – for ***obydwa*** `both', ***kilkoro*** – for ***kilka*** `several', ***kilkanaścioro*** – for ***kilkanaście*** `a dozen or so', ***naścioro*** – for ***naście*** `a dozen or so', ***kilkadziesięcioro*** – for ***kilkadziesiąt*** `a few dozen'.

It is easily seen that the traditional Polish name for these units, "liczebniki zbiorowe" (collective numerals), parallel to the Russian one (собирательные числительные), does not correspond to their meaning. Contrary to Russian, in contemporary Polish the meaning of the forms considered does not contain the additional (in relation to regular cardinal numerals) component `all', causing the noun phrase to be considered as a homogenous group (see Mel'čuk, 1982, 1985). Such a component is present in the lexemes ***oba*** / ***obydwa*** (including the series ***oboje*** / ***obydwoje***) `both' as opposed to ***dwa*** (including the series ***dwoje***) `two'. Therefore, in order to avoid that awkward term, we will call the forms under discussion simply *pięcioro* forms. Their meaning will be shown below.

The central problem of our paper is the interpretation of the case-differentiated *pięcioro* forms: whether they should be treated as separate lexemes inflected for case or included into more complicated numeral lexemes (of the type ***pięć***), with the form differentiated according to the case and gender contrast. The first solution (the traditional one) was also accepted by Laskowski (1984, 1998). The second solution was proposed by me over thirty years ago (Saloni, 1976) and included in many later articles (Saloni 1977, 1992, 2003) and handbooks (Saloni & Świdziński, 1987; Saloni et al., 2007), but it was not discussed separately. Such a discussion is the main purpose of this paper.

2 The Situation in Russian

Analogous forms in Russian correspond only partially with their counterparts in Polish. The first similarity concerns their inflection for case. Let us take parallel examples of the forms corresponding to the numbers 2 and 5:

Nom	*дв**о**е*	*п**я**теро*
Gen	*дво**и**х*	*пятер**ы**х*
Dat	*дво**и**м*	*пятер**ы**м*
Acc	*дв**о**е*	*п**я**теро*
Inst	*дво**и**ми*	*пятер**ы**ми*
Loc	*дво**и**х*	*пятер**ы**х*

In Russian tradition these series of forms are treated as the lexemes ***двое*** and ***пятеро***, inflected for case, and separate from the cardinal numerals ***два*** and ***пять***. Analogous series of forms differentiated by the value of case exist for the numbers 2, 3 (***трое***, cf. ***три***), 4 (*четверо*, cf. ***четыре***), 5, 6 (***шестеро***, cf. ***шесть***), 7 (***семеро***, cf. ***семь***), 8 (***восьмеро***, cf. ***восемь***), 9 (***девятеро***, cf. ***девять***), 10 (***десятеро***, cf. ***десять***) – and only for them.

Thus the set of forms in question is in Russian distinctly smaller than in Polish. First of all, we should notice the lack of a set of forms of the type ***двое*** / ***пятеро*** for the lexeme ***оба*** `both'. It can be seen as a natural consequence of their meaning, which includes the additional component `all', characteristic of collective numerals (собирательные числительные) and visible in such phrases as *трое солдат*, as opposed to *три солдата* (both roughly translatable as `three soldiers'), although the semantic relations between such phrases are much more complicated (see Mel'čuk, 1985). The irregularity of the meaning was the basic reason for Mel'čuk's decision that collective numerals are separate lexemes in Russian.

The interpretation of analogous numeral forms in Russian is not unanimous. Bogusławski (1966) proposed to include forms of the type *двое* into the class of homogeneous numeral lexemes of the type ***два***. The strongest argument for interpreting forms of the type *двое* co-occuring with *plurale tantum* nouns as belonging to cardinal numeral lexemes is their semantic identity with other forms of these lexemes joined with forms of nouns inflected for number. Such a solution is for Zaliznjak (1967: 87) obvious. The form *двое* in the construction *двое саней* `two sleighs' (*сани* ` sleigh' being grammatically plural) is required by the noun forms in the same way as the form *два* in the construction *два мальчика* `two boys' is re-

quired by the noun *мальчик,* or the form *две* in the construction *две девушки* `two girls' by the noun *девушка.*

3 Basic Properties of *pięcioro* Forms in Polish

A similar grammatical contrast, and more distinctive, appears in Polish.

3.1 Semantics

It is evident that the majority of Polish numerals (cardinal numerals) inflect not only for case, but also for gender. If we want to translate into Polish the English sentences *I see five boys* and *I see five girls*, we must use distinct forms of the same lexeme ***pięć***:

(1) (a) *Widzę* ***pięciu*** *chłopców.*
(b) *Widzę* ***pięć*** *dziewczyn.*

The opposition is the same as for forms of adjectives and verbs joined with noun forms of different genders[1]:

(2) (a) *To* ***był dobry*** *chłopiec.* `This was a good boy.' (masculine animate or virile: *M1*)
(b) *To* ***była dobra*** *dziewczyna.* `This was a good girl.' (feminine: *F*)
(c) *To* ***było dobre*** *dziecko.* `This was a good child.' (neuter: *N*)

However, if we want to use a form of the lexeme ***dziecko*** `child' in a sentence parallel to (1a–b), i.e., differing only by the meaning of the noun, we use neither the form *pięciu* nor the form *pięć*, but the form *pięcioro*:

(3) (a) *Widzę* ***pięcioro*** *dzieci.* `I see five children.'
(b) **Widzę* ***pięć*** *dzieci.*
(c) ***Widzę* ***pięciu*** *dzieci.*

The sentence (3c) cannot be produced by any competent speaker of Polish. Sentences of type (3b) occur in spoken Polish, but they are understood as non-standard, belonging to lower layers of contemporary Polish. Somebody using them would be treated as a non-competent user of the cultured dialect of Polish. The forms *pięciu*, *pięć*, and *pięcioro* in sentences (1a), (1b), and (3a) have the same meaning, `5'.

3.2 Morphology

The form *pięcioro* in (3a) is opposed to the numeral forms in (1a–b) on the basis of gender.

The same holds for all forms in other cases parallel to *dwoje*, i.e.:

(4)	Nom	*To jest*	*dwoje / pięcioro (*pięć) dzieci.*	`They are two/five children.'
	Gen	*Oczekuję*	*dwojga / pięciorga (*pięciu) dzieci.*	`I am awaiting two/five children.'.

[1] Following a long and important tradition, we understand grammatical gender in the noun as its syntactic property, consisting of requiring a particular form of the subordinate word. Thus grammatical gender of a noun manifests itself in the forms of the words associated with it (in the linguistic literature grammatical genders are sometimes called noun or agreement classes; in Zaliznjak (1967) they are called согласовательные классы). Aside from traditional masculine, feminine, and neuter genders in Polish, we distinguish an additional gender, which we name *co-plural* and mark with the symbol *P*. Some of these basic gender classes are subdivided. As is accepted in contemporary Polish linguistics, the masculine gender is divided into three subgenders on the basis of the accusative (singular and plural) forms of adjectives governed by the nouns (see Mańczak 1956). The subdivision of the genders *N* and *P* will be discussed further.

Dat	*Ufam*	*dwojgu / pięciorgu (*pięciu) dzieciom.*`I trust two/five children.'
Acc	*Lubię*	*dwoje / pięcioro (*pięć) dzieci.* `I like two/five children.'
Inst	*Interesuję się*	*dwojgiem / pięciorgiem dzieci (*pięcioma dziećmi).* `I am interested in two/five children.'
Loc	*Mówię o*	*dwojgu / pięciorgu (*pięciu) dzieciach.* `I like two/five children.'

We did not include here all possible variants of the constructions; in any case the occurrence of collective and regular cardinal numerals is shown. The starred forms given in parentheses may occur in nonstandard variants of Polish, more often in oblique cases (analogous forms do not occur for ***dwa***). The forms given before the parentheses are obligatory in contemporary standard Polish. The numeral forms in parentheses are used in Polish in the indicated cases with other nouns. (The vocative is syncretic with the nominative.)

Thus, this is a sufficient argument for including all *pięcioro* forms into the class of cardinal numeral lexemes (of the type ***pięć***) as special gender forms. However, this gender cannot be called simply neuter (*N*) because only some neuter nouns in Polish have the property of occurring in constructions like (4), distinguished on the basis of the contrast shown in (2). The majority of neuter nouns in Polish require other forms of numerals (regularly "cardinal"), cf. ***okno*** `window':

(5)

Nom	*To są dwa okna. / To jest pięć okien.*	`They are two/five windows.'
Gen	*Oczekuję*	*dwóch / pięciu okien.*
Dat	*Ufam*	*dwu / pięciu oknom.*
Acc	*Lubię*	*dwa / pięć okien.*
Inst	*Interesuję się*	*dwoma / pięcioma oknami.*
Loc	*Mówię o*	*dwu / pięciu oknach.*

Therefore we distinguish two neuter subgenders in Polish (only on the basis of constructions with numerals). We mark them with the symbol *N1* (for ***dziecko*** `child') and *N2* (for ***okno*** `window'). This division will be discussed later.

The second group of noun requiring *pięcioro* forms in Polish is a subset *P2* of *plurale tantum* nouns (named *co-plural* and marked with the symbol *P*). In this respect Polish is very close to Russian. Typical members of this subset are such nouns as: ***sanie*** `sleigh', ***drzwi*** `door', ***skrzypce*** `violin'. In standard Polish (and in many non-standard variants) the only possible way of joining the meaning `door' with the meaning of a given natural number is the construction with *pięcioro* forms. There is no limitation on the rank of number. If you want (in a fairytale or computer game) to express the meaning that somebody crossed 17 doors, you should say in standard Polish:

(6) *Przeszedł* ***siedemnaścioro*** *drzwi.*

Occurrences of the noun ***skrzypce*** with *pięcioro* forms are stabilized. In a string quartet there are two violins, and there are many concertos and other compositions for 2, 3, 4, and more violins. So it is easy to find proper collocations: *dwoje*, *troje*, *czworo*, *pięcioro*, *sześcioro*, *siedmioro*, and *ośmioro skrzypiec* (e.g., Andreas Vollenweider, *Koncert na ośmioro skrzypiec i orkiestrę* `Concerto for 8 violins and orchestra'). If some composer writes a composition for 18 violins, it will be called in Polish *Kompozycja na osiemnaścioro skrzypiec*. Theoretically there are no limitations on the number of counted objects. Practically, such limitations exist and we will discuss them further.

Thus, *pięcioro* forms should be treated as regular forms of the corresponding numerals for gender *N1* and *P2*. The lack of such forms for some numerals is no obstacle for such a description. In those cases the nouns of these genders co-occur with non-virile numeral forms (syncretic with genders *F* and *N2*), e.g.:

(7) (a) *Dziś w przedszkolu jest sto dzieci, choć wczoraj było tylko* ***dziewięćdziesięcioro****.*
`Today in the kindergarten there are *a hundred* children, although yesterday there were only *ninety*.'
(b) *X wprowadził tyle skrzypiec, bo miał specjalny cel:* ***siedmioro*** *skrzypiec to był symbol.*
`X introduces *so many* violins, because he had a special purpose: *seven* violins were a symbol.'

3.3 Syntax

One of main arguments against "morphological" treatment of Russian collective numerals (as belonging to regular lexemes of cardinal numbers) is, for Mel'čuk (1985), their inability to occur in complex (composite) names of numbers (of the type *two hundred thirty one*), even if semantic and pragmatic factors require using them: "inscriptions *43 сутокъ* `43 days', *74 сутокъ* `74 days', etc., cannot be pronounced orally, because **сорокъ трое* and **семьдесятъ четверо* are absolutely impossible, while *сорокъ три* and *семьдесятъ четыре* cannot join with *pluralia tantum*" (Mel'čuk 1985: 378–9, translated).

This is not the case with Polish *pięcioro* forms. They can be joined in names of numbers of higher rank. Every Pole knows the childish song:

(8) *OjciecWirgiliusz* `Father Virgilius
uczył dzieci swoje, taught his children;
a miał ich wszystkich and he had altogether
sto dwadzieścia troje. a hundred and twenty three of them.'

The only problem here is the co-occurrence of the form *troje* `3' with the "cardinal" *dwadzieścia* `20'. The difficulty is caused by the question of rules for creating such complex names of numbers, which in Polish are rather complicated and not entirely stabilized (see Gruszczyński & Saloni, 1978). However, the construction *sto dwadzieścioro troje dzieci* is not only possible, but fully acceptable (although very rare).

Such collocations, e.g., *dwadzieścioro dwoje*, *dwadzieścioro troje*, *dwadzieścioro czworo*, *trzydzieścioro dwoje*, *trzydzieścioro troje*, *trzydzieścioro czworo*, are represented on the Internet by at least 90 occurrences. They join, first of all, with the word *dzieci*; however other nouns are also widely represented there, e.g., *uczniów* `students', *turystów* `tourists', *solistów* `soloists', *maluszków* `toddlers'. These nouns are neither of gender *N1* nor *P2*, but *M1* (masculine animate or virile). Here we see an additional syntactic property of *pięcioro* forms. We will discuss it in the next section.

Special attention should be brought to names of complex numbers with the last component `1', e.g., 21, 31, etc. In such constructions containing *pięcioro* forms we use for this `1' the form *jeden* (in all genders and cases). The same applies to all names of complex numbers with the final component `1'. So we say:

(9) (a) *Było tam* ***dwudziestu jeden*** *chłopców.* `There were twenty one boys there.'
(b) *Było tam* ***dwadzieścia jeden*** *dziewczyn.* `There were twenty one girls there.'
(c) *Było tam* ***dwadzieścioro jeden*** *dzieci.* `There were twenty one children there.'
(d) *Było tam* ***dwadzieścioro jeden*** *uczniów.* `There were twenty one students there.'

There are a few dozen occurrences of this collocation on the Internet.

4 The Scope of Using *pięcioro* Forms

Using *pięcioro* forms in Polish is certainly limited. An obvious factor is coordination of their meaning with their appearance in texts: the higher the number, the lower the probability that constructions will occur with *pięcioro* forms. It is observable even in collocations with the main noun requiring them, ***dziecko***: *dwoje* / *troje* / *czworo dzieci* are obligatory, *pięcioro … dziesięcioro dzieci* relatively frequent (*pięć …*

dziesięć dzieci treated as incorrect), *jedenaścioro ... dziewiętnaścioro dzieci* rare (*jedenaście ... dziewiętnaście dzieci* tolerated), *dwadzieścioro ... dziewięćdziesięcioro dzieci* very rare[2].

Let us discuss now in detail the situations when *pięcioro* forms are used. In the first two subsections we will discuss basic (gender) rules when such a form is required by a noun form: for gender *N* and *P*. In the last subsection we will examine using *pięcioro* forms optionally with masculine animate nouns (*M1*).

4.1 Constructions with Neuter Nouns

According to general opinion, the use of *pięcioro* forms is very limited. We must agree with this opinion, although we insist on our gender interpretation of this form.

It is true that the set *N1* is very small. There are only several nouns where *pięcioro* forms are obligatory. First, there is the lexeme ***dziecko*** discussed broadly in the previous sections. Second, there are two neuter nouns denoting binary sense organs: ***oko*** `eye' and ***ucho*** `ear'. In the basic meaning the only possibility for expressing definite plurality is using *pięcioro* forms. This fact is true not only for the number `2', but also for others, e.g., for `4' only *czworo oczu* / *uszu* (in such constructions with numerals we use the archaic dual forms). However, in secondary meanings these lexical units change their grammatical properties: they have regular plural forms and join with regular (*N2*) numeral forms, e.g., there are *dwa (cztery) oka w sieci* `two (four) meshes in a net' or *dwa (cztery) ucha przy torbie* `two (four) handles on a bag'.

Third, as traditional handbooks of Polish grammar state, *pięcioro* forms join with neuter nouns denoting non-adult creatures. More precisely, they are limited to some structural group of these nouns. These are nouns ending with ***-ę*** (and having an extended stem in other forms of the singular ***-ęci-*** and plural ***-ęt-***), e.g., ***kurczę*** `chicken', ***szczenię*** `puppy' – they belong to the set *N1*, which is not very large. In the *Grammatical Dictionary of Polish* (Saloni et al., 2007), it consists of about 100 lexemes. The diminutives of *N1* nouns, e.g., ***kurczątko***, ***szczeniątko***, as well as other neuter nouns that are names of creatures, e.g., ***maleństwo*** `little one', do not belong to *N1*, but to *N2*.

A problem with this group of nouns is that their appearance is limited. They are used only in some regions of Poland (in the standard cultured dialect), but are generally understood and accepted. They do not belong – with two exceptions – to my own active vocabulary. As a theoretician I would accept them only with *pięcioro* forms. Laskowski, who descends from the region where they are actively used, wrote that they occur with these forms optionally (Laskowski, 1998:63).

Two lexemes from this group that can be used in all parts of Poland are ***zwierzę*** `animal' and ***dziewczę*** `maiden', The first belongs to fundamental biological terminology on the level of primary school, the second is pathetic in singular, but neutral and common in plural (*dziewczęta*). Both can be used in either construction: *pięcioro zwierząt* (*dziewcząt*) / *pięć zwierząt* (*dziewcząt*). Forms of the noun ***zwierzę*** are rarely used with numeral forms, but *dziewczęta*, *dziewcząt*, etc., are easily and neutrally used with non-collective forms. The relatively popular German musical production *Das Dreimäderlhaus* `The House of Three Maidens' played in Poland under the title *Domek trzech dziewcząt*. I have known this title for a half century, but noticed that there is something suspicious in the formulation only when I started to analyze how Polish numeral forms are used.

In modern Polish some archaic phraseological expressions with *pięcioro* forms can also be met, e.g., *dziesięcioro przykazań* `ten commandments' or *doktor obojga praw* `Doctor of both laws', although the nouns ***przykazanie*** and ***prawo*** in contemporary Polish belong to the gender *N2*[3].

4.2 Constructions with Co-Plural Nouns (P*luralia Tantum*)

The co-plural nouns in Polish (as in Russian) are divided into two main subsets, corresponding to the division of nouns inflected for number. The main gender contrast in the plural in Polish is between mascu-

[2] There are characteristic lacunae in the sequence of lexemes corresponding to them in Polish dictionaries (see Kucała, 1976).

[3] There is also an archaic collocations with a *pięcioro* form with a feminine noun: *ludzie płci obojga* `persons of both sexes'.

line animate (*M1*) nouns and other nouns. In accusative plural we have two possibilities of gender agreement:

(10) (a) *Widzę* ***dobrych*** *chłopców.* `I see good boys.' (*M1*)
(b) *Widzę* ***dobre*** *dziewczyny.* `I see good girls.' (*F*, also *N*, *M2*, *M3*)

The same contrast is observable between groups of co-plural nouns:

(11) (a) *Widzę* ***dobrych*** *rodziców.* `I see good parents.' (*P1*)
(b) *Widzę* ***dobre*** *skrzypce.* `I see good violin(s).' (*P–1*)

Nouns occurring in contexts like (11a) we will mark *P1*, analogously to *M1*. The complement of the subset *P1* will be temporarily marked in a parallel fashion as *P–1*; it will be further subdivided.

As we have shown, some nouns of the subset *P–1* occur with *pięcioro* forms. To our examples ***sanie*** ` sleigh', ***drzwi*** `door', ***skrzypce*** `violin' we may add further: ***grabie*** `rake', ***widły*** `pitchfork', ***cymbały*** `dulcimer', and also ***nożyce*** `shears', ***nożyczki*** `scissors'.

Two last examples are characteristic: in English, as well as in Russian and in many other languages, they have *plurale tantum* equivalents. This is the consequence of their physical structure – as paired objects, consisting of two mutually symmetric parts. And for such objects there is another way of expressing specified plurality: *dwie pary nożyc (nożyczek)* `two pairs of shears (scissors)'. This is a surrogate, lexical means of expressing plurality, which exists in various languages.

This is a rule for other *pluralia tantum* (*P*) nouns in Polish, especially referring to sets of two separate symmetric objects, which are destined for symmetric parts of the human body, e.g., ***buty*** `shoes' (with all possible kinds and varieties), ***rękawiczki*** `gloves', ***skarpetki*** `socks', ***pończochy*** `stockings, ***narty*** `skis', etc. For all of them there are corresponding regular two-number nouns, e.g., ***but***, ***rękawiczka***, ***skarpetka***, ***pończocha***, ***narta*** (the same equivalence exists in English and other languages). The noun phrase *para butów* (*rękawiczek*, *skarpetek*, *pończoch*, *nart*) is synonymous with the lexeme ***buty*** denoting an ordered set of two objects: and combining such a phrase with a numeral is the only way of expressing definite plurality of such objects. These constructions are created according to the general rules for joining forms of numerals and nouns in Polish (see Saloni, 1977), e.g.:

(12) (a) *Mam* ***parę*** *butów.* `I have **a pair of** shoes.'
(b) *Mam* ***dwie pary*** *butów.* `I have **two pairs of** shoes.'
(c) *Mam* ***pięć par*** *butów.* `I have **five pairs of** shoes.'

It is impossible to add a numeral form immediately to such a noun form. Therefore, we divide the class of non-virile *pluralia tantum* (*P–1*) into to subclasses: *P2* (which occur with numeral forms, i.e., with the *pięcioro* form) and *P3* (which do not occur with any numeral form).

The core of the set *P2* is shown above. It consists of names of distinctively formed material objects that do not have a paired structure (although they are often symmetrical), like violin, rake, or pitchfork. The most interesting case is the noun ***drzwi*** `door'. A door is usually asymmetrical, however many types of doors are not only symmetrical, but consist of two separate parts, approximately mutually symmetrical. Such types of doors have Polish *plurale tantum* nouns, e.g., ***wrota***, ***wierzeje***, ***podwoje***, which cannot co-occur with numeral forms; they belong to the set *P2*.

The border between sets *P2* and *P3* is not distinct. The words ***nożyce*** and ***nożyczki*** are examples of nouns for which definite plurality may be expressed in both ways: immediately with numerals (i.e., with *pięcioro* forms) or in descriptive constructions with the noun ***para*** `pair'. This border group is very broad; it contains the names of clothes (and similar objects) having a symmetrical structure: ***spodnie*** `trousers', ***majtki*** `underpants', ***kalesony*** `long johns', ***okulary*** `glasses', etc. (its dual structure is seen in the *plurale tantum* form, also in English equivalents), which cannot be joined with numeral forms – they belong to the set *P2*. If we count such objects, we use descriptive constructions much more often, e.g. *pięć par*

spodni, *dwie pary majtek* (*okularów*). The constructions with numeral forms occur much more rarely, e.g., *pięcioro spodni*, *dwoje majtek* (*okularów*).

We may distinguish two more groups of nouns belonging to the set *P–1*: the names of events: ***urodziny*** `birthday', ***imieniny*** `name day', ***zajęcia*** `classes (a unit of timetable in universities and other schools)' (there is analogous group of *pluralia tantum* in Russian), and names of substances: ***perfumy*** `perfume', ***fusy*** `dregs', ***pomyje*** `swill'. Counting such objects is rather atypical (see Kucała, 1976), but – for some of them – quite possible. It happens that somebody participated twice in another person's birthday or had classes with two different groups of students and wants to communicate this fact. He has two possibilities: to use a descriptive way of communicating or to use a *pięcioro* form, e.g.:

(13) *Byłem na* ***dwojgu*** *urodzinach.* `I was at **two** birthday parties.'

This possibility is rather theoretical, but one can find on the Internet occasional examples of such constructions. There is no other possibility for combining forms of these nouns with numeral forms. The same concerns names of substances that can be joined with numerals in secondary meanings, e.g.:

(14) *Lubi tych* ***dwoje*** *perfum.* `She likes these **two** perfumes.'

In the *Grammatical Dictionary of Polish* (Saloni et al., 2007), we marked such nouns *P2*, although the correctness of sentences like (14) may be called into question.

Virile co-plural nouns *P1* in Polish also co-occur with *pięcioro* forms. There are two subsets of them.

The first (and much more numerous) subset consists of names of (married) couples, e.g., ***rodzice*** `parents', ***dziadkowie*** `grandparents, literally: grandfathers' ***narzeczeni*** `betrothed (couple)', etc. We should pay attention to a specific subset of this group: lexemes derived from regular *M1* nouns with the suffix *-(o)stwo*[4], e.g., ***państwo*** `Mr. and Mrs.', ***profesorostwo*** `professor with his wife', ***wujostwo*** `uncle (*wuj*) with his wife', ***kuzynostwo*** `cousin (*kuzyn*) with his wife'. This derivational class is productive. It is quite possible to create occasionally such nouns, although their use is limited. They are treated as obsolete, official, not common. There are only over a hundred such nouns in Polish dictionaries (e.g., Saloni & al., 2007). All refer only to objects that theoretically can be counted, but it is not possible to express the meaning of plurality by a construction with such a noun, e.g., **pięć*/*pięcioro profesorostw.* However, they can occur with numeral forms–of the lexemes ***oboje*** and ***obydwoje*** `both', e.g.:

(15) (a) *Zaproś* ***oboje*** *dziadków..* `Ask both grandparents.'
(a) *Zaproś* ***oboje*** *profesorostwa.* `Ask both: the professor and his wife.'

This is a special use of the numeral forms: we do not count sets denoted by the noun, but objects belonging to them.

The second subset of *P1* nouns includes several names of sets of people, but not ordered pairs, e.g., ***państwo***[5] `ladies and gentlemen', ***rodzeństwo*** `siblings'. They share some of the properties of the first subset: it is impossible to use them in the plural sense, e.g., they cannot join with forms of quantifying adjectives, e.g. ***każdy*** `each', ***żaden*** `neither', ***wszystek*** `all' (cf. Zaliznjak & Paducheva, 1974:32). However, they can join with numeral forms that serve to count objects belonging to the set denoted by the noun. We have found, for example, the following examples on the Internet:

---

[4] The *P1* nouns with this suffix (also belonging to the second subset), although syntactically plural (they occur with plural forms of adjectives and verbs), have morphological exponents of singular forms: their desinences are the same for the singular of two-number names. Therefore, we have systematic homonymy of forms of nouns with this suffix, e.g. ***braterstwo*** 1. *N2* `brotherhood', 2. *P1* `brother with his wife'.

[5] There are three lexemes in Polish having the basic form ***państwo***. Besides the two *pluralia tantum* discussed in this section, there is a regular two-number lexeme with the meaning `state; country' (in constructions with numerals: *dwa państwa*, *pięć państw* etc).

(16) (a) *Za pomyłkę* ***troje Państwa*** *przepraszam.* `I apologize to three of you for my mistake.'
(b) *Miał* ***czworo rodzeństwa***. `He had four siblings.'

In both cases the set should be understood as containing persons of both sexes.

In the *Grammatical Dictionary of Polish* (Saloni et al., 2007), we decided to describe both those groups as virile *pluralia tantum*, against Zaliznjak's proposal to describe Russian lexemes of the type *родители*, *девчата* as belonging to (potentially) two-number defective nouns, having no singular forms (Zaliznjak, 1967:99).

4.3 Additional Rules – Constructions with Masculine Animate (*M1*) Nouns

There is one additional situation when *pięcioro* forms are used: with masculine animate (*M1*) nouns. It is well known that basic numeral forms joining with forms of this gender are different. However in contemporary Polish it is possible to use the following constructions:

(17) (a) *Na egzaminie było* ***pięciu*** *studentów.* `There were five students at the exam.'
(b) *Na egzaminie było* ***pięcioro*** *studentów.*

The meaning of sentence (17b) is evidently richer than that of (17a): the experienced competent speaker of Polish says (or understands) that the group of five students was sexually mixed: it contained at least one man and one woman. Sentence (17a) does not include such an element of meaning. It may be understood in two ways: either that there were five male students or that there were five students whose sex is not of our concern. Its meaning can be specified in broader contexts, e.g.:

(18) (a) *Na egzaminie było* ***pięciu*** *studentów i cztery studentki.*
. `There were five male students and four female students at the exam.'
(b) *Na egzaminie było* ***pięciu*** *studentów. I same kobiety.*
. `There were five students at the exam. And only women.'

It is clear that only M1 nouns that do not contain the semantic element `male' occur in constructions with *pięcioro* forms. Nouns having such semantic element cannot be used with these forms, e.g., \**pięcioro chłopców* literally would mean `five boys of both sexes'. Generally, in Polish *M1* nouns can refer to persons independently of their sex.

This way of using *pięcioro* forms is correct only with *M1* nouns. There are feminine nouns denoting persons regardless of sex, e.g., ***osoba*** `person', ***istota*** `being'. However, constructions like ****pięcioro osób***, which are sometimes encountered, especially in spoken language, are treated by educated Poles as incorrect, not acceptable.

5 Conclusion

The first meaning of *pięcioro* forms is the same as for other forms corresponding to numbers (in our example: `5'). The basic distribution of these forms, contrasted with other forms of the same meaning, is motivated by the category of gender, and therefore they should be included in the same homogeneous lexemes. The full paradigm of the typical numeral ***pięć*** `5' is the following (for the gender *M1* we do not include only the additional forms used for mixed groups):

	M1	*M–1, N2, or F*	*N1, P1, or P3*
Nom	*pięciu*	*pięć*	*pięcioro*
Gen	*pięciu*		*pięciorga*
Dat	*pięciu*		*pięciorgu*

Acc	*pięciu*	*pięć*	*pięcioro*
Inst	*pięciu/pięcioma*		*pięciorgiem*
Loc	*pięciu*		*pięciorgu*
Voc	*pięciu*	*pięć*	*pięcioro*

In addition to their gender functions they have a special additional meaning: they refer to sexually mixed groups of persons.

Unfortunately, *pięcioro* forms do not belong to those frequently used in Polish. However, they are obligatory with the noun ***dziecko*** `child', which both is very frequent and easily occurs with numerals. This noun is the core of a gender class. And this property suffices to regard *pięcioro* forms as a gender variant of cardinal numerals.

References

Bogusławski Andrzej. 1966. *Semantyczne pojęcia liczebnika i jego morfologia w języku rosyjskim*. Wrocław: Ossolineum.

Gruszczyński Włodzimierz & Zygmunt Saloni. 1978. Składnia liczebników we współczesnym języku polskim. *Studia gramatyczne* II, Wrocław:17–42.

Kucała, Marian. 1976. O rodzaju gramatycznym w języku polskim. [In:] R. Laskowski (ed.), *Kategorie gramatyczne grup imiennych we współczesnym języku polskim,* Wrocław: 31-42.

Mańczak, Witold. 1956. Ile rodzajów jest w polskim? *Język Polski*, XXXVI:116-121.

Mel'čuk, Igor A. 1982: И.А. Мельчук, Лично-количественные («собирательные») числительные в русском языке. *Russian Linguistics*, 6:307-334.

Mel'čuk, Igor A. 1985: И.А.Мельчук, *Поверхностный синтаксис русских числовых выражений.* Wiener Slawistischer Almanach, Sonderband 16, Wien.

Laskowski, Roman. 1984. Zagadnienia ogólne morfologii. [In:] R. Grzegorczykowi, R. Laskowski, H. Wróbel (eds.), Morfologia. Warszawa (2nd ed. 1998).

Saloni, Zygmunt. 1976. Kategoria rodzaju we współczesnym języku polskim. [In:] R. Laskowski (ed.), *Kategorie gramatyczne grup imiennych we współczesnym języku polskim,* Wrocław:43-78.

Saloni, Zygmunt. 1977. Kategorie gramatyczne liczebników we współczesnym języku polskim. *Studia gramatyczne*, Wrocław: 145-173.

Saloni, Zygmunt. 1992. New Remarks on Polish Numerals. [In:] A. Clas (ed.), *Le mot, les mots, les bons mots*. Montréal 1992:169-182.

Saloni, Zygmunt. 2003. The Problem of the Syntactic Head in Polish and Russian Constructions with Numerals. (Sidenotes to Igor Mel'čuk's Book *The Surface Syntax of Russian Numeral Expressions*). [In:] *Proceedings of the First International Conference on Meaning–Text Theory*. Paris: 193–200.

Saloni, Zygmunt & Marek Świdziński. 1987. *Składnia współczesnego języka polskiego*. 3. ed..Warszawa.

Saloni, Zygmunt, Włodzimierz Gruszczyński, Marcin Woliński & Robert Wołosz. 2007. *Słownik gramatyczny języka polskiego*. Warszawa: Wiedza Powszechna.

Zaliznjak, Andrej A. 1967: А.А.Зализняк, *Русское именное словоизменение*. Moscow.

Zaliznjak, Andrej A., & Elena V. Paducheva. 1974: А.А.Зализняк, Е.В.Падучева, О контекстной синонимии единственного и множественного числа существительных. *Информационные вопросы семиотики, лингвистики и автоматического перевода* (Moscow), 4:30–35.

Towards a semantically oriented selection of the values of $Oper_1$. The case of *golpe* 'blow' in Spanish

Begoña Sanromán Vilas
PL 59 (Unioninkatu 38 B)
FI-00014 University of Helsinki
Finland
begona.sanroman@helsinki.fi

Abstract

Contrary to the general claim that values of $Oper_1$ of a given argument are merely tools to integrate the substantive argument into a syntactic structure without entailing any extra meaning, this paper seeks to demonstrate that selection from the various values of $Oper_1$ is based on meaning. In other words, the semantic links the verbal values establish with the substantive argument on the one hand, and with the free verbal counterpart on the other. The hypothesis of this study is that each of the values of $Oper_1$ emphasizes a particular (set of) semantic component(s) of the definition of the argument and rather than being arbitrary, the selection of the (set of) component(s) coincides with the part of the meaning the value shares with its free verbal counterpart. The Spanish noun *golpe* 'blow' and its verbal values for $Oper_1$ will be used to test this hypothesis.

1 Introduction

One of the most neutral ways of expressing the meaning '*golpe'* ('blow') in Spanish in a verbal form is the verbalization of *golpe*, *golpear* 'to blow'. Alternatively, a collocation can be formed by combining *golpe* with a value of the lexical function (LF) $Oper_1$, which takes the base of the collocation, *golpe*, as the direct object and the first semantic actant of *golpe*, the agent, as the subject; e.g., *dar un golpe*/*golpes* 'to give a blow/blows'. Values of $Oper_1$, or support verbs, in the context of a collocation are semantically empty,[1] which means that they are not selected by their own lexical meaning. Consequently, if the only semantic contribution support verbs add to the collocation is to incorporate the predicate expressed by the noun into the time, it can be claimed that the expressions *golpear* and *dar un golpe/golpes* are equivalent. Using the comparison made by Bolshakov & Gelbukh (1998), the meaning of *dar* in *dar un golpe* is the same as that of the suffix *–ar* in *golpear*.

Nevertheless, what occurs with the various values of $Oper_1$? From a theoretical viewpoint, it can be said that *administrar* 'to administer', *lanzar* 'to throw' or *soltar* 'to let go of', when combined with *golpe*, have the same meaning as *dar* in *dar un golpe*, so, they should be considered as equivalents. However, examples in (1) and (2) contradict the previous statement.

(1) El Estado establece su escala, el orden de valores, mediante la variación en el número de *golpes administrados* al criminal o por el número de meses o años que se le quitan. [2]
'The Government establishes its scale, the order of values, by means of the variation in the number of blows given to the criminal or by the number of months or years that are taken away from him.'

[1] More information about the emptiness of these verbs will be offered later.
[2] The examples used in this paper are taken from the *Corpus de referencia del español actual* (CREA) <http://www.rae.es>, several Spanish dictionaries and other online sources. Some proper names have been changed and examples abbreviated.

(2) a. El Estado establece su escala, el orden de valores, mediante la variación en el número de *golpes dados* al criminal o por el número de meses o años que se le quitan.
b. El Estado establece su escala […] mediante […] el número de *golpes* **lanzados* al criminal…
c. El Estado establece su escala […] mediante […] el número de *golpes* **soltados* al criminal…

The support verb *administrar* within the collocation *administrar golpes* in (1) can be replaced by *dar* (2a) but not by *lanzar* (2b) or *soltar* (2c). The reason that *dar* can substitute *administrar* is that its aspectual information is broad enough to cover the specific one carried by *administrar*. The use of *administrar* indicates that the blows are distributed at regulated time intervals, and they are calculated in order to obtain a particular result, similarly to what happens when *administrar* co-occurs with medicaments (*Se les administró 120 mg. de magnesio por día durante 3 meses* 'They were administered 120 mg. of magnesium per day for 3 months'). Neither *lanzar* nor *soltar* can substitute *administrar* because they do not share the same aspectual information and they also impose other restrictions: in the meaning of *lanzar* a certain distance is assumed between the agent and the receiver of the blow; *soltar* includes instead a component referring to the loss of control by the agent.

A collocational dictionary such as the *Diccionario de colocaciones del español* (DiCE)[3] (Alonso Ramos, 2005), or the *Lexique actif du français* (LAF) (Mel'čuk & Polguère, 2007), classifies collocatives of a given lemma in the form of lists supplying semantic information by means of a gloss (Alonso Ramos, 2006). This semantic information is crucial in facilitating dictionary usage. Nevertheless, as collocatives are organized according to the different types of LFs, values of the same LF are provided with the same gloss remaining undifferentiated.

The main goal of this paper is to show the differences among a group of values of the LF $Oper_1$ applied to the argument *golpe* as a first step to devise a means to add this information to lexical entries of dictionaries such as the *Dictionnaire explicatif et combinatoire du français contemporain* (DEC) (Mel'čuk et al., 1984-1999), DiCE and LAF. Based on some preliminary observations, the hypothesis of this study is that each support verb emphasizes a semantic component of the definition of *golpe* and the selection of this component is not arbitrary, but coincides with that part of the meaning the support verb shares with its correlated full verb. For instance, *estampar* ('to stamp') as a support verb of *golpe (golpe de X a Y en Z con W* 'X's blow to Y in Y's part of the body Z with the instrument W'), in (3),

(3) Le *estampó* un golpe en la cara.
'He dealt him a blow in the face' (lit. 'to stamp')

focuses on 'the resounding effect of the blow when X's hand touches violently Y'. At the same time, the resounding effect due to the strength of the touch, emphasized by the support verb *estampar*, correlates with the component 'to leave a mark on a surface…' of *estampar* ('to stamp, print') as a full verb (4):

(4) Resolvió publicarlas en la misma imprenta […] que acababa de *estampar* la primera edición…
'He decided to publish them in the same printing house that has just printed the first edition…'

As can be seen in through examples (3) and (4), the coincidence between semantic components does not mean that the emphasized component of *golpe* and that of the full verb must be identical. As the support verb and its free verbal counterpart are two different lexical units (LUs), what the support verb retains from the other is some kind of semantic link, very often metaphorically or metonymically transformed in order to be adapted to the new situation.

If the hypothesis can be verified, then it is possible to understand why nouns from different semantic fields share the same support verbs and, on the contrary, why nouns included in the same semantic field

[3] The websites at <http://www.dicesp.com> and <http://dicesp.cesga.es>, include a demo of DiCE with ten lexical entries of emotion nouns.

do not combine with the same support verbs. The verbal collocative *estampar* co-occurs not only with *golpe* 'blow', but also with *beso* 'kiss', *firma* 'signature' or *pie* 'foot'. The four nouns pertaining to different semantic fields: nouns referring to hostile contacts such as *golpe*, nouns denoting friendly contacts such as *beso*, nouns alluding to written marks representing a person such as *firma* or nouns referring to traces or instruments which can leave a trace such as a *pie*. On the contrary, out of the following nouns, *paliza* 'an undetermined number of blows in a unit of time', *puntapié* 'kick given with the end of the foot', *bofetada* 'slap', and *coz* 'blow given by an animal with its leg', only *bofetada* can select *estampar* as support verb. Apart from the audible impact on Y's body, *estampar* focuses on other components such as 'the mark must be made with X's hand' (this excludes *puntapié* and *coz*) and 'there is a single blow at once' (this excludes also *paliza*). Even if the hypothesis is proved to be true, we must still explain what happens to other types of support verbs such as those which have no independent verbal counterparts, e.g., *propinar* or *infligir*, or those which are full verbs, such as *pegar* or *sacudir*. Several comments will be made in this respect throughout this paper. This study forms part of more general research aimed at establishing the paraphrase among collocations with different values of the LF $Oper_1$ + noun and the verbalization of the noun.

This research is carried out within the theoretical and methodological framework of the Explanatory and Combinatorial Lexicology (Mel'čuk et al., 1995). In this framework, collocations are described by means of LFs (Wanner, 1996, among others). Several Spanish dictionaries have been used in order to analyze the data (see bibliography).

This paper is organized into four sections. After this introduction, section 2 provides a classification of the values of $Oper_1$ applied to the noun *golpe*. In section 3, an in-depth analysis is made of support verbs sharing semantic components with their free verbal counterparts. Section 4 summarizes the paper and draws some conclusions.

2 Classification of the values of $Oper_1$ (*golpe*)

The Spanish noun *golpe* is a vocable which consists of several LUs. It can refer to a physical contact between two entities (*Dio varios golpes con el martillo* 'He gave several blows with the hammer'), the effect of that contact (*Tenía un golpe en la cabeza* 'He had a blow in the head'), an unpleasant emotion (*La muerte de Valentina fue un duro golpe para la familia* 'Valentina's death was a hard blow for the family'), a witticism ([talking about a film] *Tiene algunos golpes para morirse de risa* 'It contains some witticisms to laugh at death'), a robbery (*Una duda daba vueltas en torno del asalto: por qué dieron el golpe de día* 'There was still a doubt about the robbery: why was it perpetrated during daylight'), etc., as well as to function as a collocative itself, specifically as a value of the LF Sing ('a portion of'), e.g., *Una noche se despierta con un golpe de tos* 'One night he wakes up because of a sudden coughing fit'. The sense considered in this paper denotes physical contact and can be defined as follows:

(5) *golpe de X a Y en Z con W* = 'acto por medio del cual el individuo X, que está en movimiento, entra en contacto con la entidad Y, X tocando Y en la parte Z (de Y) con una parte del cuerpo de X o un instrumento W'
'X's blow to Y in Z with Z' = 'act by means of which individual X, who is in movement, comes into contact with entity Y, X touching Y in Y's part Z with X's body part or with an instrument W'

Ex.: *Le ha dado unos golpes y ha vuelto a funcionar* [the computer] 'He has given some blows and it has worked again'; *Prefiere que lo saluden con un golpe en la espalda* 'He prefers to be greeted with a blow on the back'; *Le propinó un golpe en la cabeza al agente*, 'He gave him a blow on the head', etc.

From the above definition it can be deduced that *golpe* can cover a large number of situations because it does not contain specific information. It can refer either to an intentional or non-intentional contact, a

calculated or spontaneous one; it can stand for friendly, hostile or neutral physical contacts; the same noun is valid for slow and fast blows, periodical or iterative ones, with different levels of intensity, given with the hand, fist or whatever object, in any part of Y's body if it is an individual; the blow can be also expected or unexpected as well as directed at the target or delivered at random, and it can produce different kinds of effects.

There are plenty of verbs which can combine with *golpe*. In this paper, 34 verbs are presented. Out of these, *golpear* is the only one which does so on a paradigmatic level (*Pedro golpea en el rostro con su puño derecho a Marcos* 'Pedro punches Marcos in the face with his right fist'). Although *golpear* is the verbalization of *golpe* (V_0), it cannot be used for all the meanings of *golpe*, for instance, 'to commit a robbery' or 'to say witticisms'. On the contrary, all the senses in the vocable *golpear* include *golpe*. In general, this verb, *X golpea Y en Z con W* 'X causes one or several blows to Y in Z with W' can be labelled as a contact verb; however, taking into account that the contact implies also motion because at least one of the participating entities has to be in movement to be able to enter into contact with the other, *golpear* is also a motion verb.

The remaining verbs are values of the LF $Oper_1$ applied to the noun *golpe*. In other words, they are support verbs which take *golpe* as the DO and the first semantic actant of *golpe* as the subject. From a syntagmatic viewpoint, they are semantically $empty_2$ verbs, which means that they are not selected by their lexical meaning; their function is to provide the noun with temporal, modal, aspectual information, etc. (Alonso Ramos, 2004: 87). From a paradigmatic viewpoint, they can be semantically $empty_1$ in the sense that they have an abstract meaning including only generic components which characterize the semantic class of the verb (Alonso Ramos, 2004:85). This is the case of *dar*, the most frequent support verb (Koike, 2001: 84),[4] which expresses 'action'. According to some authors (Reuther, 1996; Alonso Ramos, 2004), not all support verbs are semantically $empty_1$, sometimes they maintain semantic links with other senses of the corresponding full verbs. This assumption leads Alonso Ramos (2004: 91-93) to develop a scale to classify support verbs according to their more or less $empty_1$ character. Thus, five support verb classes can be distinguished in the scale: 1) pure support verbs which only have taxonomic meaning; e.g., *dar* 'to give', *hacer* 'to do/make', *tener* 'to have'; 2) support verbs with similar semantic components to their free verbal counterparts; e.g., *gozar (de salud)* 'to enjoy good health', which preserves the component 'pleasant'; 3) support verbs without free verbal counterpart, e.g., *perpetrar (un crimen)* 'to perpetrate a crime'; 4) support verbs with a homonym verbal counterpart, e.g., *librar (una batalla)* 'to fight a battle', lit. 'to free'; and 5) semantically full support verbs, e.g., *decir (una mentira)* 'to tell a lie'.

Taking Alonso Ramos' scale of emptiness in support verbs as a starting point, we have classified 33 values of the LF Oper1 applied to *golpe* in the following way (see Table 1).

SUPPORT VERBS WITH FREE VERBAL COUNTERPARTS			SUPPORT VERBS WITHOUT FREE VERBAL COUNTERPARTS	
pure support verbs	sharing some semantic components with the counterpart	full verbs	only for the semantic field 'golpe'	other fields
dar	administrar, asestar, atinar, calcar, descargar, disparar, encajar, encasquetar, endilgar, endosar, espetar, estampar, lanzar, largar, mandar, meter, plantar, soltar, tirar, proporcionar, ensartar, aplicar, suministrar	arrear, atizar, cascar, endiñar, pegar, sacudir, zumbar	propinar	infligir

Table 1. Classification of 33 values of $Oper_1$ (*golpe*) according to their $empty_1$ character

[4] See Alonso Ramos (1997) for a more detailed description of *dar*.

We have considered two major blocks: one which includes all the support verbs with free verbal counterparts and a second which includes support verbs without free verbal counterparts. In the first block, there are three groups: pure support verbs with *dar* as the unique example (3% of the total amount of verbs studied); support verbs sharing some components with the corresponding counterparts, where 70% of the verbs are included; and support verbs with the same meaning as the counterparts or full verbs (representing 21% of the totality). This classification differs from that of Alonso Ramos in that it does not include a group for support verbs with homonymic free verbal counterparts. Sometimes the limits to establishing the existence of a semantic link between the support verb and the free verbal counterpart are very fine. Some dictionaries include meanings that others do not and the researcher has to take a decision in each particular case. In the sample of verbs under study, the support verb *plantar* has finally been considered semantically connected through a metaphorical link to its free counterpart 'to put in the land a seed or plant in order to let it grow'; however, the border between this solution or that assuming a relation of homonymy is very close.

Even though the pure support verb *dar* is not discussed in detail here, it can be presented as the most neutral value of $Oper_1$ applied to *golpe* and, consequently, that which, compared to the others, can replace *golpear* in most contexts.[5] Its meaning is so general that it refers only to 'action'. On the contrary, the collocatives *arrear*, *atizar*, *cascar*, *endiñar*, *pegar*, *sacudir* and *zumbar* (different forms with the meaning 'to blow, beat') are located at the other end of the scale of emptiness; in other words they have the same meaning as their free counterparts. Therefore, one way to explain why collocations in (6a) and (7a) can replace their free counterparts in (6b) and (7b) and vice versa must be that the meaning of the collocative is subsumed by the meaning of the substantive predicate.

(6) a. Y cogió un palo, abrió la puerta y ya le iba a *arrear un golpe* al perro…
'And he took a stick, opened the door and he was already giving a blow to the dog…'
b. ¿Y si va y me *arrea* con el candelabro…?
'And if he beats me with the candelabra…?'

(7) a. El policía saca su porra reglamentaria y le *sacude un golpe* en la coronilla.
'The policeman takes his truncheon and gives him a huge blow on the crown of his head'
b. Tropieza con la barredora y ésta levanta una escoba y le *sacude* en la cabeza.
'He trips over the road sweeper and she raises a broom and beats him over the head'

From a diachronic point of view, it would be interesting to know whether *arrear*, *atizar*, *cascar*, etc., became full verbs with the meaning 'to blow' due to their frequent combination with nouns denoting 'blow' or they were in fact collocatives co-occurring with nouns such as *golpe* and because of that they absorbed in some way the meaning of the noun to end up being full verbs. According to Moliner (DUE, under *pegar*), the LU *pegar* as a collocative comes earlier than the LU as a full verb.[6]

The second block includes support verbs without free verbal counterparts, that is to say, they function only as verbal collocatives. Here we distinguish verbs which co-occur only with nouns of the semantic field of 'golpe', in which *propinar* is included (3% of the total sample), from those which combine with nouns of other semantic fields: the case of *infligir* (3%), selected not only by nouns like *golpe*, but also nouns denoting intense physical or moral suffering (*tortura* 'torture', *dolor* 'pain'), nouns referring to punishments (*condena* 'sentence' *sanción* 'sanction'), mistreatment (*ataque* 'attack', *traición* 'treason'), failure in sports competitions (*derrota* 'defeat', *goleada* ''lots of goals'), etc. *Infligir* does not in fact operate in totally different semantic fields, but rather in a single broader one, where nouns alluding to damages, within nouns like *golpe,* are included.

[5] Although *dar un golpe/golpes* and *golpear* are semantically equivalent in a broad sense, many other differences can be found between both expressions which prevent them from being interchangeable in all contexts (see Sanromán Vilas, forthcoming).
[6] "De la acep. 'pegar', aplicar una cosa a otra, se pasaría a la de 'aplicar un golpe' y, después, a la de 'golpear'".

In order to explain what collocatives without a free counterpart contribute to the collocation they form part of, a possible solution would be to generalize the common characteristics to all the nouns in co-occurrence with them. When the collocative combines only with a particular field as in the case of *propinar*, a look back at its etymology may throw some light on the question. In this vocable, there is an old meaning of *propinar*, 'to give tips', which relates it to other transference verbs such as *administrar* 'to administer', *dar* 'to give', *suministrar* 'to supply', etc.

To end this section, a tentative classification of the support verbs is presented in Table 2 following partly previous semantic-syntactic classifications of Spanish verbs such as Vázquez et al. (2000) and ADESSE (García-Miguel et al., 2005). This classification is based on the meaning of the free verbal counterpart or, if no such counterpart exists, on a feature common to all the nouns from other semantic fields co-occurring with the verbal collocative or on an old meaning of the verb from which it may have inherited certain features. Each verb is labelled with a semantic tag taking as a starting point *golpear*, classified at the beginning of this section primarily as a contact verb and secondly as a motion verb.

MOTION VERBS *X* **Oper$_1$(*golpe*)** *un golpe/golpes a Y (en Z) con W*						
CONTROL	DISPLACEMENT	ORIENTATION	TRANSFERENCE	CONTACT	LOCALIZATION	MODIFICATION
largar*, soltar	disparar*, lanzar*, largar*, mandar*, tirar*	asestar, atinar, disparar*, lanzar*, tirar*	administrar, dar, endilgar, endosar, mandar*, propinar, proporcionar, suministrar, descargar*	aplicar, arrear, atizar, calcar, endiñar, ensartar*, estampar*, pegar, sacudir, zumbar	descargar*, encajar, encasquetar, ensartar*, espetar, meter, plantar	cascar, estampar*, infligir

Table 2. Semantic classification of 33 values of Oper$_1$ (*golpe*)

Table 2 shows that support verbs co-occurring with *golpe* have the same basic propositional form and are included in the group of motion verbs, for entity X has to initiate a movement to enter into contact with entity Y. Within motion verbs seven other categories are distinguished: 1) CONTROL VERBS, which focus on the loss of control of X over something which starts to move, e.g., *soltar*; 2) DISPLACEMENT VERBS, indicating the movement of an entity from its original place, e.g., *lanzar*; 3) ORIENTATION VERBS, which denote a movement towards the target point Y, e.g., *asestar*; 4) TRANSFERENCE VERBS, referring to the fact that X moves something from the place where X is to the place where Y is, e.g., *dar*; 5) CONTACT VERBS, focusing on the fact that entity X touches entity Y, e.g., *pegar*; 6) LOCALIZATION VERBS, which indicate the way the movement arrives entity Y and stays there, e.g., *plantar*; and 7) MODIFICATION VERBS, denoting the effect of the movement upon Y, e.g., *cascar*. When a verb has an asterisk, it means that it takes part in two categories; for instance, *mandar* is at the same time a transference and displacement verb.

3 Support verbs sharing semantic components with the free verbal counterparts

In this section a selection of support verbs sharing semantic components with their free verbal counterparts is presented[7] using a pattern which contains the following information:

a) *Semantic tag*: Here the verb is classified according to a general semantic tag (Table 2 is followed).
b) *Co-occurrence*: This part contains a sample of nouns (from different semantic fields than *golpe*) which combines with the verb as a value of Oper$_1$.

[7] *Dar* as a pure support verb, full support verbs as well as support verbs without free verbal counterparts are excluded here.

c) *Usage label*: This heading denotes those verbs used in a particular register (formal or colloquial), as part of jargon or slang or even showing the speaker's attitude towards the enunciation, for instance, emotive, pejorative, etc.
d) *Spanish variety*: This includes a label indicating the geographical variety of Spanish.[8]
e) *Gloss*: This section is provided with a description of the meaning of the support verb in combination with *golpe*.
f) *Semantic links*: This section explains what the collocative inherits from its free verbal counterpart.[9]
g) *Emphasized component(s) of the definition of 'golpe'*: Here, as well as indicating the part of the meaning of *golpe* singled out, there is a reference to how the stress is created.[10]
h) *Characteristics*: We add here additional information such as specific syntactic characteristics, restricted lexical co-occurrence, e.g., adjectives or adverbs selected neither by the noun *golpe* nor by the verb independently but by the whole phrase support verb + noun, etc.
i) *Examples*: Some examples from a corpus are included here to illustrate usage.

This pattern can be taken as a starting point to specify distinctions among values of $Oper_1$. Below, a selection of support verbs, including various semantic tags are analyzed.

administrar lit. 'to administer';
SEMANTIC TAG: transference verb; CO-OCCURRENCE: medicaments and sacraments; USAGE LABEL: formal.
GLOSS: 'X, having power over Y or having a recognized knowledge in a particular area, X causes that X, X's body part or an instrument W, comes into contact with Y touching Y's part Z a certain number of times at regulated time intervals'.[11]
SEMANTIC LINK WITH THE FREE VERBAL COUNTERPART: From the full verb ('X administers Y', 'X has the power or the knowledge to organize, decide and distribute something in a particular way'), the collocative retains the aspectual information (it takes place at regulated time intervals).
EMPHASIZED COMPONENTS OF THE DEFINITION OF 'GOLPE' (shown in (5)):
- the attributes of the individual X: 'X having power over Y or recognized knowledge in an area';
- the act is intentional and calculated (it is done in a particular way to obtain an intended result).
CHARACTERISTICS AND EXAMPLES: In (8), the police represent the entity with power; the authority in (9) comes from the fact that the extract is located in a first aid manual. In (9) the mode of action is regulated by series of five blows.
(8) *[dos policías] le administraron golpes en las orejas y patadas para hacerlo caer.* '[two policemen] gave him blows in the ears and kicks to make him fall.
(9) *[primeros auxilios en caso de atragantamiento de un niño de corta edad] Los golpes dorsales consisten en administrar golpes en la zona situada entre los omoplatos… Se recomienda hacer series de cinco golpes dorsales…* '[first aid for a small child who is choking] Dorsal blows entail administering blows in the area located between scapulas… Giving a series of five dorsal blows is recommended…

atinar lit. 'to hit on';
SEMANTIC TAG: orientation verb; CO-OCCURRENCE: value of $Oper_1$ only for the semantic field of *golpe*.
GLOSS: 'X intending to come into contact with Y, X aims at Y; X not being certain that X (X's body part or the instrument W) touches Y's part Z, X comes into contact with Y by chance'.[12]

[8] Points (c) and (d) are not included if the support verb belongs to a neutral, non-specialised register, and to standard Spanish.
[9] For instance, how the original meaning of the free verbal counterpart is transformed by means of metaphorical or metonymical links; if there is no free counterpart, a disappeared meaning within the vocable or something shared by other nouns combined with the collocative can serve the same function. The collocative frequently inherits the lexical aspect from its counterpart.
[10] It must be remembered that the definition of *golpe* (5) consists of neutral semantic components, so singling out any of these components means specifying them expressly.
[11] 'X, que tiene poder sobre Y o que tiene conocimientos reconocidos en un área, X causa que X, una parte del cuerpo de X o un instrumento W, entre en contacto con Y, X tocando la parte Z de Y un número de veces en intervalos de tiempo regulados'.

SEMANTIC LINK WITH THE FREE VERBAL COUNTERPART: It retains the idea of the orientation from the realization verb *atinar**1*** ('X atina Y en Z con W') 'X reaches Y in the place Z by orienting the weapon W' and the meaning 'by chance' from *atinar**2*** ('X atina con Y') 'X finds Y by chance'. *Asestar (un golpe)* is another support verb which inherits the meaning 'orientation' from a realization verb (Real(*cañón* 'cannon')).[13]
EMPHASIZED COMPONENTS OF THE DEFINITION OF 'GOLPE':
- the act is specified as intentional (even if the result is achieved by chance) and orientated towards Y.
CHARACTERISTICS AND EXAMPLES: Examples in a negative form are frequent (10); *golpe* is often combined with *certero* 'accurate' (11); *atinar* can be also found subordinated to the verb *tratar* 'to try' (11).[14]
(10) *No atina un golpe ni por casualidad* 'He was not able to hit a blow, not even by chance'
(11) *[...] tratando de atinar un golpe más certero*... 'trying to hit a more exact blow…'

disparar lit. 'to shoot';
SEMANTIC TAG: displacement and orientation verb; CO-OCCURRENCE: value of $Oper_1$ only for the semantic field of *golpe*; USAGE LABEL: often used in boxing jargon; SPANISH VARIETY: According to DRAE, *disparar* is used colloquially in Cuba with the meaning 'to beat a person or an animal', e.g., *Le disparó un golpe*.
GLOSS: 'X using a special technique, X comes into contact with Y; X, X's body part or the instrument W, touching Y's part Z with a directed, precise, and fast movement, X causing a damage in Y'.[15]
SEMANTIC LINK WITH THE FREE VERBAL COUNTERPART: From the meaning of *disparar* ('X [individual] sends Y [projectile] to Z [individual] in W [Z's body part] with K [weapon]') another one has been originated by metaphorically associating 'weapon' with 'fist', and more specifically, the movement and the trajectory of the projectile with the technique of using the fist;
EMPHASIZED COMPONENTS OF THE DEFINITION OF 'GOLPE':
- the act is specified as a calculated one; orientated towards Y and made at high speed;
- the result of the act causes a strong impact on Y.
COMMENTARIES AND EXAMPLES: Taking, for instance, *bala* 'bullet' or *pistola* 'pistol' as arguments of a LF, *disparar* 'to shoot' is a realization verb, that is, a value of the LF Real or Labreal; however, this is not the case in *disparar un golpe*, which is only a more precise paraphrase of *dar un golpe* or *golpear*. When *disparar* is a value of Real(*bala*), the whole collocation *disparar una bala* can be replaced by the full verb *disparar* 'to shoot a weapon'; but the full verb *disparar* cannot substitute collocations in (12) and (13). *Disparar* could be considered a special value of Real(*golpe*) in boxing jargon, but not in colloquial Spanish, the sense represented here. [16]
(12) *El baterista se sintió tan frustrado [...]que disparó un golpe a un vaso y el vidrio traspasó su mano*.
'The drummer was so upset […] that he launched a blow at a glass and the glass went through his hand'
(13) *[...] lo cual no quiere decir que le haya clavado una mirada furiosa y disparado un golpe de puño*...
'[…] which does not mean that I had stared at him and dealt a blow with my fist…'

estampar lit. 'to stamp';
SEMANTIC TAG: contact, modification; CO-OCCURRENCE: *firma* 'signature', *autógrafo* 'autograph'; *huella* 'footprint', *mano* 'hand'; *beso* 'kiss', etc.; USAGE LABEL: colloquial.
GLOSS: 'X causes that X comes into contact with Y, X's hand touching violently one or more times Y's body part (generally the face), X causing a resounding effect and often leaving a physical mark on Y'.[17]

---

[12] 'X teniendo la intención de entrar en contacto con Y, X se orienta hacia Y; X no teniendo la certeza de que X (la parte del cuerpo de X o el instrumento W) toque la parte Z de Y; X entra en contacto con Y casualmente'.
[13] See in commentaries, under *disparar*, why these verbs are here considered values of $Oper_1$.
[14] Out of eight Spanish monolingual dictionaries, none registers *atinar un golpe/golpes*.
[15] 'X usando una técnica especial, X entra en contacto con Y, X, una parte del cuerpo de X o el instrucmento W, tocando la parte Z de Y con un movimiento rápido, preciso y directo, X causando un daño en Y'
[16] *Disparar* is also used as a value of complex LFs such as Incep/CausPred Plus (*Los precios se disparan* 'prices rise up', *Algo dispara las ventas* 'something increases sales')
[17] 'X causa que X entre en contacto con Y, X tocando de manera violenta una o más veces una parte del cuerpo de Y (generalmente, la cara), X causando un efecto sonoro y a menudo dejando una marca física en Y'.

SEMANTIC LINK WITH THE FREE VERBAL COUNTERPART: From the full verb *estampar* 'to stamp, to print', 'X leaves the mark Y [letters, drawings] on Z [paper or other materials] by pressing an instrument with the hand', the collocative *estampar* takes the components: strength (by pressing), focus on the result (the mark and the sound) and aspectual information (generally action only occurs once in a unit of time (semelfactive), e.g., *estampar un golpe*, *un par de besos*, etc.; if it is a repeated action, the number of blows is not undetermined (*Le estampó dos sonoras bofetadas*, **Le estampó una paliza*)).
EMPHASIZED COMPONENTS OF THE DEFINITION OF 'GOLPE':
- the intensity of the action;
- the result of the action: it leaves a mark (print, hint) as well as a resounding effect;
- X's part of the body: the contact is done with X's hand.
EXAMPLES:
(14) *Le estampó un golpe en la cabeza…* 'He dealt a blow to the head'
(15) *Cerró el puño y estampó un golpe sobre mi rostro*. 'He closed the fist and dealt a blow to my face'

plantar lit. 'to plant';
SEMANTIC TAG: localization verb; CO-OCCURRENCE: *beso* 'kiss', *base* 'basis', etc.; USAGE LABEL: colloquial.
GLOSS: 'X causes that X comes into contact with Y, X touching Y's face with a sudden and fast movement; Y not expecting to be touched by X, Y has no time to react'.[18]
SEMANTIC LINK WITH THE FREE VERBAL COUNTERPARTS: The basic meaning of the vocable *plantar* is 'X puts Y (a plant, a seed, etc.) in Z by introducing Y partially in the earth'. The link with this definition is probably that X is active and Y passive. Other LUs within the vocable *plantar* focus on the component 'X having a commitment to Y, X abandons Y' or 'to put something or somebody in a place in a sudden and abrupt way'; from the first of them, the collocative *plantar* retains the idea that X's action is unexpected by Y; from the latter the association is with the suddenness of the movement.
EMPHASIZED COMPONENTS OF THE DEFINITION OF 'GOLPE':
- the act is a spontaneous one, made with a fast movement, and it is unexpected by Y;
- Y's body part Z is generally the face.
CHARACTERISTICS AND EXAMPLES: (16) and (17) shows that the act is unexpected, sudden and in the face.
(16) *Lo toma por sorpresa y le planta un golpe en la cara que lo deja perplejo y en silencio total* 'He takes him by surprise and dealt him a blow to the face, leaving him puzzled and in total silence'
(17) *Repentinamente le planta un golpe de frente entre la nariz y la boca* 'Suddenly he dealt him a blow straight between the nose and the mouth'

soltar lit. 'to let go of';
SEMANTIC TAG: control verb; CO-OCCURRENCE: communication nouns (*insulto* 'insult', *comentario* 'comment', *amenaza* 'threat', etc); *estornudo* 'sneeze', *risotada* 'guffaw'; USAGE LABEL: colloquial.
GLOSS: 'X having avoid to express an emotion, X causes that X comes into contact with Y, X, X's body part or an instrument W, touching Y's part Z with a sudden movement, X feeling that X has expressed the emotion'. The intensity of the blow is proportional to the intensity of the emotion;[19] e.g., in *soltar* the intensity is neutral, in *descargar* (a close meaning), is positive.
SEMANTIC LINK WITH THE FREE VERBAL COUNTERPARTS: From the meaning of the full verb *soltar*, 'X causes that Y, which has been tied, stops being tied', the support verb retains the idea that the released object has been repressed. This aspect distinguishes *soltar* from *largar un golpe*, despite their proximity in the propositional meaning.[20] The movement in *largar* comes out without previous repression.
EMPHASIZED COMPONENTS OF THE DEFINITION OF 'GOLPE':

[18] 'X causa que X entre en contacto con Y, X tocando la cara de Y con un movimiento brusco y rápido, Y no esperando ser tocado por X, Y no tiene tiempo para reaccionar'.
[19] 'X que ha estado evitando expresar una emoción, X causa que X entre en contacto con Y, X, una parte del cuerpo de X o un instrumento W, tocando la parte Z de Y con un movimiento brusco, X expresando la emoción. La intensidad del golpe es proporcional a la intensidad de la emoción'
[20] There is a communicative difference between them: the use of *largar* implies a negative attitude from the speaker's side.

- X stops controlling the emotion; as a result, X feels somewhat released;
- The movement leading to the contact is sudden and abrupt.

CHARACTERISTICS AND EXAMPLES: The movement is not calculated in relation to the target (18); there is no intention to cause damage, only to be released (19).

(18) *Es importantísimo hacer algo más que soltar golpes a diestro y siniestro*. 'It is very important to do something more than land blows left, right and centre'

(19) *Nunca supe si irme o quedarme. Si soltar un golpe cariñoso o…* 'I never knew whether to leave or to stay. Whether to land a warm blow or…'

A description of each support verb in the dictionary could help users to select the suitable collocative in a particular context. Elements contained in the previous descriptions such as "usage label" or "Spanish variety" can be introduced directly in dictionaries; however, elements such as the emphasized components of the definition of the argument should be standardized prior to their inclusion in the dictionary. In this sense, we propose formalizing the semantic components of the definition of *golpe* using the concept of *semantic dimension*, understood as a "set of two or three mutually exclusive values" (Mel'čuk & Wanner, 1996: 216) which provides easy access to the semantic components of a definition. As an example, we could propose semantic dimensions for the definition of *golpe* such as emotionality, directionality, or noticeability, among others, with the following values:

> emotionality 'contact expressing an emotion' ={'friendly' (+), 'hostile' (–), 'emotionality-neutral'}
> expectation 'Y, (un)expecting the contact' ={'expected' (+), 'unexpected' (–), 'expectation-neutral'},
> noticeability 'the result of the contact can be noticed' ={'noticeable' (+), 'noticeability-neutral'}, etc.

The above dimensions have an unmarked value in *golpe*'s decomposition because they correspond to trivial semantic characteristics of the noun (see note 10). However, when a collocative singles out one of these potential components of the noun, it becomes marked. For instance, if *golpe* selects *plantar*, it means that the blow was unexpected by Y; when *estampar* is chosen, the result of the blow leaves a sign, etc. In this sense, our proposal would be to include semantic dimensions of *golpe* (e.g., 'unexpected' and 'noticeable') in its lexicographic entry beside the collocative (e.g., *plantar* and *estampar*, respectively).

4 Conclusions

The aim of this paper was to test the hypothesis that selection among the various values of $Oper_1$ is based on the semantic links verbal values establish with the substantive argument on the one hand, and with the free verbal counterpart on the other. In this sense, we have attempted to show that different values of $Oper_1$ applied to *golpe* emphasize a particular (set of) semantic component(s) of the definition of the noun *golpe* and the selection of the (set of) component(s) is not arbitrary but normally has some relation to the part of the meaning the value shares with its free verbal counterpart. One possible way to complete this study in order to be able to apply the results to practical lexicography would be to formalize the semantic components of the definition of *golpe* using semantic dimensions and to include the emphasized dimension(s) by each collocative in the entry of *golpe* beside the collocative.

Acknowledgements

This paper was written within the framework of the research project FFI2008-06479-C02-01/FILO (Ministerio de Educación y Ciencia, Spain). I would like to thank two anonymous reviewers whose critical comments helped to improve the manuscript.

References

Alonso Ramos, Margarita. 1997. Coocurrencia léxica y descripcion lexicográfica del verbo dar: hacia un tratamiento de los verbos soportes. *Zeitschrift für Romanische Philologie*, 113(3):380-417.

Alonso Ramos, Margarita. 2004. *Las construcciones con verbo de apoyo*. Madrid, Visor Libros.

Alonso Ramos, Margarita. 2005. Semantic Description of Collocations in a Lexical Database. In F. Kiefer et al. (eds.), *Papers in Computacional Lexicography COMPLEX 2005*, Budapest: Linguistics Institute and Hungarian Academy of Sciences, 17-27.

Alonso Ramos, Margarita. 2006. Glosas para las colocaciones en el Diccionario de Colocaciones del Español. In M. Alonso Ramos, *Diccionario y Fraseología*, Coruña: Universidade da Coruña, 59-88.

Bolshakov, Igor A., & Alexander Gelbukh. 1998. Lexical Functions in Spanish. In *Proc. Simposium Internacional de Computación*, 11-13 novimenbre 1998, México D.F.: CIC, 383-395. In Internet [consulted 05/03/2009]: <http://www.data.cicling.org/CV/Publications/1998/CIC-98-Lexical-Functions.htm>

García-Miguel, José M., Lourdes Costas, & Susana Martínez. 2005. Diátesis verbales y esquemas construccionales. Verbos, clases semánticas y esquemas sintáctico-semánticos en el proyecto ADESSE. In G. Wotjak & J. Cuartero Otal (eds.), *Entre semántica léxica, teoría del léxico y sintaxis*, Frankfurt am Main: Peter Lang, 373-384.

Koike, Koike, Kazumi. 2001. *Colocaciones léxicas en el español actual. Estudio formal y léxico-semántico*. Alcalá de Henares, Universidad de Alcalá/Takushoku Unviersity.

Mel'čuk, Igor A. et al. 1984-1999. *Dictionnaire explicatif et combinatoire du français contemporain*. *Recherches lexicosémantiquesI-IV*. Montreal, Les Presses de L'Université de Montréal.

Mel'čuk, Igor A., André Clas, & Alain Polguère. 1995. *Introduction a la lexicologie explicative et combinatoire*. Louvain-la-Neuve, Duculot.

Mel'čuk, Igor A., & Alain Polguère. 2007. *Lexique actif du français. L'apprentissage du vocabulaire fondé sur 20000 dérivations sémantiques et collocations du français*. Louvain-la-Neuve, De Boeck.

Mel'čuk, Igor A., & Leo Wanner. 1996. Lexical Functions and Lexical Inheritance for Emotion Lexemes in German. In: L. Wanner (ed.), 209- 278.

Reuther, Tilmann. 1996. On Dictionary Entries for Support Verbs: The Cases of Russian vesti, provodit' and proizvodit'. In: L. Wanner (ed.), 180- 208.

Sanromán Vilas, Begoña (forthcoming). Colocaciones verbo-nominales y correlatos verbales independientes. Análisis tentativo de las diferencias.

Vázquez, Glòria, Ana Fernández, & M. Antònia Martí. 2000. *Clasificación verbal. Alternancias de diátesis*, Lleida, Universitat de Lleida.

Wanner, Leo (ed.). 1996. *Lexical Functions in Lexicography and Natural Language Processing*. Amsterdam/ Philadelphia, John Benjamins.

Spanish dictionaries and corpora

ADESSE: *Alternancias de diátesis y esquemas sintáctico-semánticos del español*, <http://webs.uvigo.es/adesse/>

CREA: RAE, *Corpus de referencia del español actual*, <http://www.rae.es>,

DEA: Seco, Manuel, Olimpia Andrés, & Gabino Ramos. 1999. *Diccionario del español actual*. Madrid, Aguilar.

DELE: Alvar Ezquerra, Manuel (dir.) 2006. *Diccionario para la enseñanza de la lengua española. Español para extranjeros*. Universidad de Alcalá/Barcelona: Vox_Bibliograf.

DRAE: RAE. 2001. *Diccionario de la lengua española*. (22ª ed.). Madrid, Espasa Calpe (in Internet [consulted 05/03/2009] <http://buscon.rae.es/draeI/>).

DSLE: Gutiérrez Cuadrado, Juan (dir.) 2006. *Diccionario Salamanca de la lengua española*. Madrid, Santillana.

DUE : Moliner, María. 2007. *Diccionario de uso del español*. (3ª ed.; also 1ª and 2ª eds.). Madrid, Gredos.

LEMA: Battaner Arias, Paz (dir.). 2001. *LEMA. Diccionario de la lengua española*. Barcelona, Spes Editorial.

PRÁCTICO: Bosque, I. (dir.) 2006. *Diccionario combinatorio práctico del español contemporáneo*. Madrid, SM.

REDES: Bosque, Ignacio (dir.). 2004. *REDES. Diccionario combinatorio del español contemporáneo*. Madrid, SM.

Russian Botanical Terms: Towards Their Lexicographic Description

Alexei Shmelev
Moscow Pedagogical State University
M. Pirogovskaia 1
Moscow 119991, Russia
shmelev.alexei@gmail.com

Elena Shmeleva
Vinogradov Institute of Russian Language
Volkhonka 18/2
Moscow 119019, Russia
eshkind@mail.ru

Abstract

The paper discusses the Russian names of plants and their parts as they are used in everyday language and in two scientific disciplines: biology and crop husbandry. In many Russian texts (e.g., in different types of "gardener's handbooks") these three terminological systems interact in quite an intricate way. The explications of botanical terms in existing Russian explanatory dictionaries are inadequate in many respects. The paper proposes a theoretical basis for the lexicographic description of the terms under consideration and also illustrates its principles with several examples. Analysis of the use of botanical terms in Russian texts along with linguistic experiments have shown that everyday (standard) word usage based on a *naïve taxonomy* has little to do with biological nomenclature. The taxonomy of crop science is closer to standard usage; however, some differences exist. A complete lexicographic description should take account of the usage of "botanical words" in different terminological systems (providing appropriate labels where necessary).

1 Introduction

The linguistic semantics of 1960–1970s paid little attention to terms for natural kinds. Thus, Apresjan (1967:19) suggested that the semantic code of such a noun might be its number in the relevant dictionary. In line with this, at the early stage of development of the Meaning ↔ Text Theory, it showed little interest in the semantics of natural kinds.

Now the situation has radically changed. It comes as no surprise that Apresjan et al. have devoted a special section (2007:91–93) to explications of the meaning of Russian names for various kinds of fruit (*фрукты*).

In what follows, we will discuss the Russian "botanical words", that is, names of plants and their parts. We will proceed from the general principle that the lexical entry should be oriented to the standard word usage based on naïve concepts; however, a complete lexicographic description should take account of the usage of "botanical words" in various genres including "scientific" usage (providing appropriate labels where necessary). The more so, many of the "naïve" speakers are inclined to believe that the scientific model of the world and the corresponding word usage is more "correct" and pattern their own usage on the scientific one. However, one must be cautious about the impact of scientific or quasi-scientific ideas on the "naïve" conceptualization of the world. The explicit statements of language speakers should be taken with a grain of salt. Anna Wierzbicka comments about a similar situation, "if a scientifically oriented informant claims that 'for him', tigers **are** cats, my reaction would be rather to mistrust him as an informant (because he cannot distinguish colloquial English from scientific terminology), than solemnly to write down that in his 'lect' tigers are a kind of cats." (Wierzbicka, 1985:42) What counts is how the speakers use language, not their explicit "metalinguistic" statements.

In dividing up a polysemous word (a vocable, according to the terminology of the Meaning ↔ Text Theory) into separate meanings (lexemes, according to the terminology of the Meaning ↔ Text Theory),

one should also take into account, among other things, grammatical behavior in different types of usage. If a word behaves differently in different types of usage, one may conclude that they belong to different lexical meanings.

On this basis, a dictionary entry on the noun *цветок* 'flower' may look as follows as suggested in (Šmeleva & Šmelev 2007):

(1) **цветок** 1. (pl. *цветы* or *цветки*) Colorful part of plant located on a *branch* or *stem. На ветке расцвели ярко-красные цветы*.
1a. (scientific) (pl. *цветки*) Sexual reproductive organ in plants.
2. (pl. *цветы*) Plant branch or stem with *цветы 1*. *Подарить цветы; сорвать цветок*.
3. (pl. *цветы*) Small plant growing in the ground whose visible part consists of a stem and *цветок 1* or several *цветы 1*. *На клумбе росли цветы*.
4. (pl. *цветы*) Small *indoor* (growing in earth indoors, located in a pot) plant. *В знак того, что явка провалена, на окне стоял горшок с цветком*.

Recognizing (2), (3), and (4) is made on the basis of denotation and seems self-evident.[1] As for distinguishing (1a and (1b), it is supported by differences in grammatical behavior: the genitive plural is *цветов* or *цветков* for *цветок 1*, *цветков* for *цветок 1a*, and *цветов* for *цветок 2*, *цветок 3*, and *цветок 4*. It should be added that in numeral constructions the form *цветков* is used for all meanings.

In addition, a complete lexicographic description should take account of the usage of "botanical words" in different regional varieties of language. Consider, for example, the definition of the word *урюк* given in the *Small Academic Dictionary* (Evgen'eva 1981-1984): "fruit of an apricot-tree dried with pits" (*высушенные с косточками плоды абрикоса*). This definition does not take into account the use of the word with reference to apricot-trees in Central Asia (while the standard Russian word for 'apricot-tree' is *абрикос*, exactly as a donkey may be referred to as *ишак* in Central Asia while the standard Russian word for 'donkey' is *осел*) as in an "Indispensable manual for the composition of articles for gala occasions, feuilletons for state holidays, odes, hymns, and also poems for parades" from *The Golden Calf* by Ilf & Petrov (*Literary Verse* with *Asiatic Ornamentation*):

(2) Цветет урюк под грохот дней, / Дрожит зарей кишлак. / А средь арыков и аллей / Идет гулять ишак.
'Under daily din the *ooryuk* [apricot-tree] blooms, / Like dawn the *keeshlak* [village] flares, / While midst *aryks* [canals] and alley glooms / The *eeshak* [ass] forth he dares.
(Authorized translation from the Russian by Charles Malamuth)

Needless to say we do not mean that the descriptions of "botanical words" should be extended so as to include all regional varieties of language; nor do we aim at dialects as well standard Russian. The point is that the use of the word *урюк* we are interested in is known to most speakers of Russian and is widely used in speech and written texts as in the following example from *Cancer Ward* by Alexander Solzhenitsyn:

(3) В первое утро творения — кто ж способен поступать благорассудно? Все планы ломая, придумал Олег непутёвое: сейчас же, по раннему утру, ехать в Старый город смотреть цветущий урюк.
'The first morning of creation — who can act rationally on such a day? Oleg discarded all his plans. Instead he conceived the mad scheme of going to the Old Town immediately, while it was still early morning, to look at a flowering apricot tree.'

One can also find many examples of the use of *урюк* with reference to apricot-trees in the *National Corpus of the Russian Language*. This leads us to conclusion that a lexicographic description should take

[1] One may observe that the semantic structure of the English word *flower* does not have the meaning similar to the fourth meaning of the word *цветок*: if the visible part of an indoor plant does not have a colorful part that catches one's eye and grows on a stalk or twig, this plant can be only referred to as a *plant* (and never *flower*) in English.

account of the use in question as the *Small Academic Dictionary* does with regard to the words *арык* (defined as 'irrigation canal in Central Asia, Kazakhstan, Transcaucasia [в Средней Азии,Казахстане, Закавказье: оросительный канал]'), *ишак* (defined as 'ass, donkey [осел]'), and *кишлак* (defined as 'village in Uzbekistan and Tajikistan [селение в Узбекистане и Таджикистане]').

2 Plants and Their Edible Parts

Now we are in a position to discuss a fragment of Russian folk botany compared with the scientific description of the same domain of reality in order to illustrate certain problems of the relation between everyday and scientific concepts. We will deal with the names of plants and their edible parts.

2.1 Naïve usage vs. special purpose usage

This domain of reality was chosen not by chance. Its specific character lies in the fact that it is studied at least in two scientific disciplines: biology and crop husbandry.[2] Each of these disciplines has its own terminological system, and the models of the world of these disciplines (both of which differ from the naive model) are not the same (although the world model of crop science is admittedly much closer to the naive model than the scientific biological model). In many Russian texts (e.g. in different types of "gardener's handbooks") these three models of the world interact in quite an intricate way.

In order to assess the differences between everyday speech, crop science and biology, we need only turn to the Russian biology schoolbook for grades 6-7. Different types of fruit are described there as follows:

(4) Поскольку плоды очень многообразны, для разных плодов существуют специальные названия. Сочный односемянный плод называют *костянкой*. Наружный слой околоплодника у него – тонкая кожица, средний – сочная мякоть, внутренний – деревянистый, образует косточку (вишня, слива, персик). Сочный многосемянный плод – *ягода*. Наружная часть околоплодника состоит из тонкой кожицы (помидор, виноград, смородина). *Яблоко* – сочный многосемянный плод (яблоня, груша, рябина). В его образовании принимает участие не только завязь, но и другие части цветка. *Зерновка* – сухой односемянный плод с тонким пленчатым околоплодником, сросшимся или плотно прилегающим к семенной кожуре (рожь, пшеница, кукуруза). *Семянка* – сухой односемянный плод с кожистым околоплодником, не срастающимся с семенной кожурой (одуванчик, подсолнечник). *Орех* – сухой односемянный плод с деревянистым околоплодником (орешник, липа). *Желудь* – сухой односемянный плод, в отличие от ореха околоплодник кожистый. *Коробочка* – сухой многосемянный плод, открывающийся крышечкой (белена), дырочками (мак) или створками (тюльпан). *Боб* и *стручок* – сухие многосемянные плоды. Но у стручка между створками есть перегородка, на которой и располагаются семена (капуста, редька), а у боба такой перегородки нет.
'Given the great diversity of fruit, special names exist for different types of fruit. A fleshy one-seeded fruit is called *костянка* 'stone fruit.' The outer layer of the seed vessel is a thin membrane, the middle layer consists of fleshy pulp, and the ligneous inner layer forms the pit (cherry, plum, peach). A fleshy many-seeded fruit is called *ягода* 'berry.' The outer part of the seed vessel consists of a thin membrane (tomato, pear, currant). *Яблоко* 'pome' is a fleshy many-seeded fruit (apple, pear, rowanberry). It forms not only from the ovary but also from other parts of the flower. *Зерновка* 'grain' is a dry one-seeded fruit with a thin membranous seed vessel that is fused with or adpressed to the seed hull (rye, wheat, corn). *Семянка* 'achene' is a dry one-seeded fruit with a leathery seed vessel that does not fuse with the seed hull (dandelion, sunflower). *Орех* 'nut' is a dry one-seeded fruit with a ligneous seed vessel (hazel, linden). *Желудь* 'acorn' is a dry one-seeded fruit with a leathery seed vessel (in contrast to the nut). *Коробочка* 'seedpod' is a dry many-seeded fruit dehiscing through a lid (henbane), holes (poppy) or valves (tulip). *Боб* 'bean' and *стручок* 'pod' are dry many-seeded fruits. A pod has a septum that bears the seeds (cabbage, radish), while a bean lacks such a septum.'

[2] Crop husbandry is primarily a branch of agriculture, but the same name refers to an applied discipline that elaborates the scientific foundations of this branch of agriculture.

If we try to understand this classification in detail, we will see that it does not mention *тыквины* (cucumber, melon, pumpkin) and *померанцы* (orange, lemon), which are close to *ягоды* 'berries'; indeed, they are sometimes referred to as *ягоды*. As for raspberries and strawberries, they are considered *сборные плоды* 'aggregate fruits'. Raspberry is *сборная костянка* 'an aggregate stone-fruit', i.e. a number of fruitlets arranged together; every fruitlet is structured in the same way as a plum or an apricot. Strawberry is *сборная семянка* ('an aggregate achene'), i.e. a number of fruitlets arranged in the same way as the seeds in a sunflower, and what we call a strawberry in everyday language is the fleshy receptacle.

Clearly, this usage has little to do with everyday word usage. In standard Russian, nobody calls pears or rowan-berries *яблоки* 'pomes; apples': *яблоки* 'apples' and pears are *фрукты* 'fruits' and rowan-berries are *ягоды* 'berries'. (Note that *ягоды* are not a kind of fruit for everyday Russian.) Aggregate fruits such as raspberry and strawberry are also *ягоды* 'berries'. At the same time, apart from the special-purpose botanical language, no one calls tomatoes, cucumbers, and oranges *ягоды* 'berries': tomatoes and cucumbers are *овощи* 'vegetables' and oranges are *фрукты* 'fruits'. Notice that the word *плод* ('fruit') in everyday Russian is rarely used to refer to a part of a plant.[3]

The difference between technical terms and everyday words is not the most important thing. It is the difference between the scientific and the naïve taxonomy that is of first importance. Placing apples, pears and rowanberries into one class, peaches, apricots and cherries into another class, and tomatoes, gooseberries and watermelons into yet another class would contrasts with the naïve classification, which would rather place together apples, pears, apricots and peaches (as *фрукты* 'fruit'), on the one hand, and rowanberries, gooseberries and perhaps cherries (as *ягоды* 'berries') on the other hand. Watermelons and tomatoes do not belong to either class: watermelons are closer to *фрукты* while tomatoes belong to yet another class of *овощи* 'vegetables'. This naïve taxonomy is reflected in the grammatical behavior of the words in question: names of berries are collective nouns (this does not hold for the generic term *ягода*) while names of fruit are countable.

The contrast between *фрукты* and *ягоды* is multidimensional. The main difference is in size: people normally hold *ягоды* with a thumb and one finger (usually index finger) while *фрукты* are normally held with a thumb and at least two fingers. To eat a *фрукт*, people usually bite off or cut off from it while *ягоды* are normally put into the mouth whole. Furthermore, many naïve speakers of Russian believe that *фрукты* grow on trees while *ягоды* grow on bushes/shrubs or herbaceous plants. This may be the reason to use the word *куст* 'bush, shrub' to refer to a rowan-tree: the phrase *куст рябины* 'rowan bush/shrub' is quite common in Russian texts. Ožegov (1972) acting as a naïve speaker has defined *фрукты* and *ягоды* as follows:

(5) ***Фрукт***. *Сочный съедобный плод какого-н. дерева.*
'Juicy edible fruit of a tree'.

(6) ***Ягода***. *Небольшой сочный плод кустарников и трав.*
'Small juicy fruit of bushes/shrubs and grasses'.

That is why we would not include the component 'growing on herbaceous plants' into the definition of *банан* 'banana' as it is suggested in (Apresjan et al. 2007:92). An average speaker of Russian believes that bananas grow on trees (or treelike plants), which may be considered a part of Russian naïve picture of the world. The *National Corpus of the Russian Language* contains a dozen of examples of the phrase *банановое дерево* 'banana tree', which can be regarded as an objective evidence for excluding the idea of 'growing on herbaceous plants' from the definition of *банан*. The more so, it also contains several examples of the phrase *банановая пальма* 'banana palm', and for an average Russian speaker *пальма* is a tree rather than a herbaceous plant (consider, e.g., the definition of *пальма* in the most of dictionaries).

[3] The rare exceptions are: (1) texts written in archaic poetical style; (2) idioms, e.g. *освящение плодов* ('blessing of fruits') at Transfiguration (in accordance with the tradition of Russian Orthodox Church); (3) word usage in texts making (partial) use of the "scientific" model of the world; (4) designations of any exotic fruit when the speaker does not know its name.

Some speakers claim that a banana palm is an herb,[4] but such claims reflect the biological nomenclature; in the casual, non-terminological usage it would be ridiculous to say *пальмы и другие травы* 'palms and other herbs' (while *пальмы и другие деревья* 'palms and other trees' would be quite correct).

True enough, the above regularity (*фрукты* grow on trees, and *ягоды* grow on bushes/shrubs or herbaceous plants) is not absolute. Ožegov (1972) not in the least embarrassed by an apparent contradiction has given the following definitions, which include reference to *ягоды* growing on trees:

(7) ***Вишня***. *Плодовое дерево с сочными съедобными темно-красными ягодами, а также ягода этого дерева.*
A fruit-tree with juicy edible dark red berries as well as a berry of this tree.

(8) ***Рябина***. *Дерево с плодами в виде пучка горьковатых оранжево-красных ягод, а также самые ягоды.*
A tree with fruits growing in a cluster of bitter orange or red berries as well as berries themselves.

(9) ***Черешня***. *Плодовое дерево, сходное с вишней, а также ягода его.*
A fruit-tree resembling a cherry-tree (*вишня*) as well as its berry.

One more difference between *фрукты* and *ягоды* is that for many (but not all!) speakers of Russian *фрукты* are always parts of cultivated plants while *ягоды* may be parts of wild plants (the phrase *лесные ягоды* 'forest berries' is commonly used while *?лесные фрукты* 'forest fruit' would sound rather strange). *Фрукты* are always edible unless poisoned by somebody while *ягоды* may be poisonous by themselves.

Because of the inconsistency between the biological nomenclature and everyday usage, the biological terminology in this area is difficult to learn and easy to forget after school. The only element of the scientific classification of fruit that almost all Russian speakers know is that a watermelon is *ягода* 'berry' from the scientific point of view. Undoubtedly, the lack of a generic term for *арбуз* 'watermelon' (as well as *дыня* 'melon') in everyday usage is a contributing factor. However, although Russian speakers know that a watermelon is a berry, the standard word usage is unaffected by this fact. It is unlikely that anybody who goes to a fruit-market *купить ягод* 'to buy some berries' will buy a watermelon (should s/he do that, it would mean that s/he had changed his/her original plans). In other words, some speakers claim that *арбуз* is *ягода*, but they do not use the word *ягода* to refer to a watermelon in everyday usage.

Apart from the lack of a generic term, everyday usage lacks a clear classification in some cases as well. Thus, some Russian speakers refer to cherries as *фрукты* 'fruits' while others refer to them as *ягоды* 'berries'. This may be because they are bigger than prototypical berries (such as currants or rowanberries) and grow on a tree. Grammatically, it may be used both as a countable noun (like *фрукты*) and as a collective noun (like *ягоды*): both *варенье из вишни* (sing.) and *варенье из вишен* (pl.) are acceptable. *Виноград* 'grapes' is only used as a collective noun; it grows on a vine rather than tree; so, it is closer to berries. Nevertheless, some speakers of Russian refer to it as *фрукты* (maybe because grapes are somewhat bigger than prototypical berries and vine differs from herbaceous plants and bushes); by the way, it may be observed that Apresjan et al. (2007) have classed *виноград* as *фрукт*. Anyway, the fact that grape is *ягода* while cherry is *костянка* 'stone-fruit' according to the biological nomenclature is unlikely to affect their naïve classification.

The inconsistency between biological terminology and everyday usage has led texts striving to reflect the scientific model of the world yet catering to the mass reader to try to explain this difference. Consider, for example, the following article from *Soviet Encyclopedia* (Sovetskij… 1985):

[4] Googling Russian *пальма* returns many examples like *банановая пальма (а точнее трава, которая просто маскируется под пальму!)* 'a banana palm (or, more precisely, a herb which masquerades as a palm)'; *Кстати, пальма не дерево, а многолетняя трава* 'As a matter of fact, palm is not a tree but a perennial herb'

(10) **´ЯБЛОКО** (ботан.), сочный, обычно многосемянный нераскрывающийся плод растений сем. розовоцветных подсем. яблоневых – груши, яблони, айвы, рябины и др. В быту Я. наз. плод яблони.
‘**ЯБЛОКО** (botan.) …fruit of… pear-trees, apple-trees, quince-trees, rowan-trees, etc. In everyday usage, Я. is the fruit of an apple-tree.’

2.2 Regular polysemy of folk “botanical” terms

Various parts of cultivated herbaceous plants function as *овощи*: roots, tubers, bulbs, leafs, leafstalks, flower stalks, fruits. They are cultivated for food: thus, there are no wild vegetables (**дикие овощи*) or poisonous vegetables (**ядовитые овощи*). In contrast to fruit and berries, which are commonly used as a dessert, *овощи* are not supposed to be used in such a way. The word *овощи* (sing. *овощ*) ‘vegetables’ has two different meanings in everyday Russian: ‘1. Part of a plant good to eat cut with salt. 2. Plants cultivated to get *овощи 1*’. Note that the *Small Academic Dictionary* (Evgen’eva 1981-1984) recognizes only one meaning for the word *овощи*, namely ‘vegetable garden fruit and greens used for food (cucumbers, carrots, tomatoes, beets, etc.)’ Ožegov (1972) suggest a similar definition: ‘vegetable garden fruit and greens used for food (cucumbers, carrots, beets, tomatoes, turnips, etc.)’. But these explications does not take into account such uses as *прополка овощей* ‘weeding vegetables’; *В огороде растут овощи* ‘Vegetables grow in the garden’; *Я люблю совмещать на грядке овощи и цветы* ‘I like to combine vegetables and flowers on one bed’ (an example from the *National Corpus of the Russian Language*). It would be wrong to treat these examples as instances of a regular polysemy that may be ignored in a dictionary, since the seemingly very similar word *фрукты* ‘fruits’ can never refer to fruit trees (nor does the word *ягоды* ‘berries’ have the meaning of ‘berry bushes’).

Most of the names of cultivated plants have the same polysemic structure: the word refers to a plant as well as to a part of this plant used by people. For example, such words as *груша* ‘pear-tree; pear’, *слива* ‘plum-tree; plum’, *вишня* ‘cherry-tree; cherry’, *абрикос* ‘apricot-tree; apricot’ denote both trees and fruits of these trees (one of the rare exceptions is the existence of two different words for the apple-tree and its fruit: *яблоня* ‘apple-tree’ and *яблоко* ‘an apple’); the words *морковь* ‘carrot’ and *свекла* ‘beet’ denote plants and their edible roots. Reference to plants is quite common in everyday usage, e. g.:

(11) Расцветали яблони и *груши*, / Поплыли туманы над рекой.
‘Apple and *pear trees* were in bloom, / The mist crept over the river.’
(*Katyusha*, lyrics by Mikhail Isakovsky)

Most Russian dictionaries take this polysemy into account by indicating ‘plant’ as the first meaning and ‘part of this plant’ as the second meaning (this is apparently due to preference for “scientific” usage in which these words denote plants, while their edible parts are indicated periphrastically: *plody abrikosa*, *korni morkovi*); consider the following entries from the *Small Academic Dictionary* (Evgen’eva 1981-1984):

(12) **АБРИКОС**…**1.** Южное плодовое дерево. **2.** Плод этого дерева оранжевого цвета с крупной косточкой
‘**АБРИКОС**… **1**. A southern fruit tree. **2**. fruit (orange in color) of this tree with a large pit’

(13) **МОРКОВЬ**…**1.** Огородное растение семейства зонтичных, овощ. **2.** *собир.* Съедобные утолщенные корни этого растения оранжевого цвета
‘**МОРКОВЬ**…**1**. A garden plant of the Umbelliferae family, a vegetable. **2**. Edible incrassate roots (orange in color) of this plant’

It may be observed that Mel’čuk (1979) suggested much more sophisticated explications taking into account countability/non-countability (thus, he distinguished four lexical meanings of the word *морковь* rather than two); however the strategy he used was essentially the same (the meaning ‘a plant…’ or

'species of plant…' comes first; the meaning 'edible product…' and 'natural unit of edible product…' are considered derived from one of the above meanings).

A good case can be made for the reverse order since the meaning of 'a kind of fruit' and 'a kind of vegetable' is the basic meaning of these terms in standard Russian while the meaning of 'a plant cultivated for this fruit/vegetable' may be considered a derived meaning. Consider the following observation made by Apresjan et al. (2007:91): "It is quite correct to say in Russian *персики, апельсины, виноград и другие южные фрукты* 'peaches, oranges, grapes, and other southern fruits' but to say *виноград, хмель и другие лианы и их плоды* 'grapevines, hops and other lianas and their fruits' would be quite ridiculous."

2.3 Crop science and everyday usage

The taxonomy of crop science is much closer to standard usage. *Плоды* ('fruit') in crop science are divided into *плоды* (apples, pears, plums, apricots, peaches), *ягоды* (gooseberries, raspberries, strawberries), *овощи* (tomatoes, cucumbers, marrow squashes), *бахчевые* культуры (watermelon, melon, pumpkin), *виноград* (grapes; a separate type). Accordingly, the following branches of crop science exist: *плодоводство* 'fruit growing' (subdivided into proper *плодоводство* and «*ягодоводство*» 'berries growing'), *бахчеводство* 'growing of melons and gourds', *овощеводство* 'vegetable growing', *виноградарство* 'viticulture'. However, the attempt to take this usage into account in an encyclopedia leads to an erosion of the "scientific" model of the world. Consider another article from *Soviet Encyclopedia* (Sovetskij… 1985):

(14) **ЯГОДНЫЕ КУЛЬТУРЫ** (ягодники), деревья, кустарники, травянистые растения, выращиваемые для получения ягод (напр., смородина, крыжовник), а также земляника, малина и др., плоды к-рых в быту неправильно наз. ягодами.
'**ЯГОДНЫЕ КУЛЬТУРЫ** (ягодники), trees, bushes, herbaceous plants cultivated for berries (e.g. currant, gooseberry bush), also strawberry, raspberry bush, etc., whose fruits are mistakenly referred to as berries in everyday usage'.

One should note that strawberries and raspberries are called *ягоды* 'berries' not only "in everyday usage", but also in the usage of crop science; hence the title *ягодные культуры (ягодники)* 'berry bushes' (in the encyclopedia there is no entry «Ягода» at all).

It seems reasonable to say that lexicographic definitions should account for the usage of "botanical words" in various types of discourse. However, it would be wrong to claim that such words as *малина* 'raspberry; raspberry bush', *земляника* 'strawberry', etc. should be defined via the notion of 'aggregate fruit' to account for the use in botanical texts. The point is that these words never refer to fruits in biological discourse. In order to refer to raspberries in a scientific botanical text, one may use a periphrastic expression *плоды малины* 'fruits of raspberry bushes'. The linguistic characteristics of words like *малина*, *земляника*, etc. are those of names for berries (the suffix *–ин-* or *–ик-* specific of names for berries, uncountability, etc.). Thus, ягода 'berry' is the only generic term for the words like *малина*, *земляника*, etc., and the structure of polysemy of the words in question may be represented along the following lines: "1. berry… 2. the plant on which this berry grows".

We would like to make one more comment concerning definitions of "botanic words". The status of different components of those definitions should be different. In particular, the components referring to size, shape, color, taste, etc. are not distinctive features of the corresponding concepts: one can easily imagine that selectionists grew *сладкие лимоны* 'sweet lemons' or *квадратные арбузы* 'square watermelons' (to make easier the process of packing). Rather, they are parts of common background knowledge about the referents of the corresponding noun: every adult native speaker of Russian would understand such expressions as *малиновый пиджак* 'crimson coat (lit. raspberry-colored coat)', *форма груши* 'pear-shaped', *величиной со сливу* 'about the size of a plum', etc. This may mean that there is a good reason to introduce these components into the definition with some additional components such as 'people think…', 'people would say…', 'usually', etc.

2.4 Tentative definitions

In this subsection, we provide tentative definitions of some terms discussed in the paper to illustrate the suggested theoretical recommendations. It should be noted that we have not included any reference to the size into the definitions of kinds of berries and fruits if this information can be deduced from the reference to the *genus proximum*.

(15) **Овощи** (sing. *овощ*) '1. Edible parts of cultivated plants growing in the ground or above the ground, usually good to eat cut with salt in order not to be hungry. 2. Plants cultivated to get *овощи 1*'.

(16) **Фрукты** (sing. *фрукт*) 'Edible fruits growing on trees or tree-like plants, too big to be put easily into the mouth, usually sweet, lightly sweet or sour, good to eat raw for pleasure; most often having seeds or a pit inside'

(17) **Ягода** 'A fruit growing on bushes/shrubs or herbaceous plants, not too big to be put easily into the mouth, often sweet, lightly sweet or sour, good to eat raw for pleasure, one after another at one time'

(18) **Абрикос** '1. Not a big *фрукт* growing on a tree in warm parts of the earth, usually oval, yellowy-orangey, with soft, sweet and juicy flesh, having a pit inside. 2. A tree on which *абрикосы 1* grow'

(19) **Урюк** '1. (uncountable) *Абрикосы 1* dried with pits. 2. *Абрикос 2* (mainly in Central Asia)'

(20) **Банан** '1. *Фрукт* growing on a tree-like plant in warm parts of the earth, usually long, slightly curved, with a thick yellow skin and whitish flesh, having no seeds inside. 2. A tree-like plant on which *бананы 1* grow'

(21) **Вишня** '1a. Edible *ягода* or small *фрукт* growing on a tree, usually round, dark red, having a pit inside. 1b. (uncountable) *Вишни 1а* as a foodstuff. 2. A tree on which *вишни 1а* grow'

(22) **Морковка** (colloquial) '1a. *Овощ*, edible root of a cultivated plant, usually long, thinner at the bottom end, orange in color. 1b. (uncountable) *Морковки 1а* as a foodstuff. 2a. (uncountable) Plants cultivated to get *морковка 1а*. 2b. A plant of *морковка 2а*'.

(23) **Рябина** '1. (uncountable) Small *ягоды* growing on a tree or a bush/shrub, usually round, red, bitter. 2. A tree or a bush/shrub on which *рябина 1* grows'

(24) **Яблоко** '1a. *Фрукт* growing on a tree, usually round or oval, red, yellow or green, whitish inside, with firm flesh, having seeds inside. 1b. (special) A fleshy many-seeded fruit of pear-trees, apple-trees, quince-trees, rowan-trees, etc.'

3 Conclusion

We hope that the presented evidence has shown that the terms for natural kinds deserve detailed linguistic analysis no less than other lexical units that used to be considered more "interesting". The more so, the terms for natural kinds throw down a challenge for a lexicographer since s/he has to be permanently drawing borderlines between what is linguistically relevant and what belongs to the extra-linguistic knowledge.

References

Apresjan, Jurij D. 1967. *Eksperimental'noe issledovanie semantki russkogo glagola*. Nauka, Moscow.

Apresjan, Jurij D., Pavel V. Djačenko, Alexandre V. Lazurskij, & Leonid L. Cinman. 2007. O komp'juternom učebnike russkogo jazyka. *Russkij jazyk v naučnom osveshčenii*, 2(14):48-112.

Evgen'eva, Anastasija P. (ed.) 1981-1984. *Slovar' russkogo jazyka*, 1-4. Russkij jazyk, Moscow.

Mel'čuk, Igor A. 1979. Countability vs. Non-countability of Nouns in Russian and Their Lexicographic Description. In *Papers from the XVth Regional Meeting of the CLS*. University of Chicago, Chicago.

Sovetskij enciklopedičeskij slovar'. 1985. Sovetskaja enciklopedija, Moscow.

Šmeleva, Elena Ja. & Aleksej D. Šmelev. 2007. Russkaja naivnaja botanika. In *Gorizonty prikladnoj lingvistiki i lingvističeskix texnologij*. DiAjPi, Partenit.

Wierzbicka, Anna. 1985. *Lexicography and Conceptual Analysis*. Karoma Publishers, Inc., Ann Arbor

Les greffes collocationnelles en espagnol

Isabel Uzcanga Vivar
Departamento de Filología Francesa
Universidad de Salamanca
uzcanga@usal.es

Araceli Gómez Fernández
Departamento de Filología y Traducción
Universidad Pablo de Olavide, Sevilla
aragomez@upo.es

Résumé

Nous étudions un cas particulier d'énoncés erronés que, suivant le terme proposé par Alain Polguère, nous appelons *greffe collocationnelle*. Nous analysons les greffes collocationnelles intralinguistiques en espagnol. Nous appliquons l'appareillage notionnel proposé par Polguère à la description de plusieurs types de greffes collocationnelles observées dans des énoncés écrits, tirés d'un corpus de mass média.

1 Introduction

Dans ce travail, nous visons l'étude d'un type particulier de collocations, qu'Alain Polguère (2007) appelle *greffes collocationnelles*, en espagnol. Selon les principes de la lexicologie explicative et combinatoire (Mel'čuk et al., 1995), une collocation est une combinaison de lexies qui est construite en fonction de contraintes bien particulières:

Elle est constituée d'une base A qui est choisie librement par le locuteur d'une langue donnée en fonction du sens qu'il veut exprimer (A), et qui contrôle la collocation au niveau fonctionnel.
Elle est constituée d'un collocatif B choisi pour exprimer un sens donné en fonction de la base, qui peut jouer deux rôles syntaxiques: a) collocatif modificateur de la base; b) collocatif de type verbe support, gouverneur syntaxique de la base

La collocation est une expression semi-idiomatique (= syntagme semi-phraséologisé). Les langues naturelles sont truffées de ce type de syntagmes, qui relèvent d'une compétence linguistique difficile à acquérir. Dans le cadre de la TST, les collocations sont décrites au moyen d'un outil formel conçu sur le modèle des fonctions mathématiques: les fonctions lexicales; plus précisément, au moyen des fonctions lexicales syntagmatiques.

Il y a longtemps que nous avions remarqué l'existence de collocations déviantes, dans des situations de parole aussi bien à l'écrit qu'à l'oral en espagnol. Ce phénomène avait éveillé notre intérêt, mais ce n'est qu'à la suite de la publication de l'article d'Alain Polguère sur les greffes collocationnelles, que nous avons disposé des notions spécifiques permettant la modélisation et l'analyse de ces collocations déviantes. D'autre part, son article nous a confirmé qu'il était nécessaire d'attribuer à ce phénomène un statut linguistique véritable.

D'après lui, «il s'agit *grosso modo* de collocations où le collocatif semble « emprunté » à une autre collocation – généralement, une collocation dont la base est sémantiquement proche de la base à laquelle le collocatif emprunté est greffé.» (Polguère, 2007: 2).

À partir des principes d'analyse proposés par Polguère nous avons analysé les données de l'espagnol.

2 Principes d'analyse

Le phénomène des greffes collocationnelles peut être considéré comme un type particulier d'interférence linguistique.

Traditionnellement, l'interférence est abordée comme un phénomène interlinguistique, où une expression appartenant à une langue L_1 est influencée par une expression appartenant à une langue autre que L_1, c'est-à-dire, à une langue L_2.

L'analogie joue aussi à l'intérieur d'une même langue : il s'agit de l'interférence intralinguistique.

Les collocations déviantes analysées dans notre corpus appartiennent à ce deuxième type d'interférence, car elles sont formées par analogie avec des collocations valides de l'espagnol.

3 Caractéristiques d'une greffe collocationnelle intralinguistique

- Elle est constituée d'au moins deux éléments lexicaux A_1 (par ex., *fe*) et B_2 (par ex., *robusta*).
- Elle fait penser à une collocation $A_2 + B_2$ bien formée (par ex., *espíritu robustecido*) ; mais elle ne l'est pas.
- Des interférences intralinguistiques se sont produites à partir de deux collocations valides de l'espagnol : une collocation valide $A_1 + B_1$ (par ex., *fe sólida, profunda, inquebrantable*), initialement visée par le locuteur, et une autre collocation également valide $A_2 + B_2$ (par ex., *espíritu robustecido*) qui s'est greffée sur la collocation initiale.

Suivant les notions proposées par Polguère, nous appellerons *cible* d'une greffe collocationnelle la collocation valide visée initialement par le locuteur, et *source* d'une greffe collocationnelle la collocation valide qui a parasité la production et qui est, donc, à l'origine de l'interférence intralinguistique, déclenchant la collocation déviante, finalement produite para le locuteur.

Le phénomène des greffes collocationnelles comprend deux types de greffes : a) les greffes qui portent sur le collocatif (dont nous nous occupons essentiellement dans ce travail) et b) les greffes qui portent sur la base de la collocation. Ces dernières sont les greffes que Polguère appelle les *greffes collocationnelles inverses*. Nous en avons des exemples dans notre corpus, et dans la plupart des cas, elles répondent à une analogie de nature phonologique.

4 Le corpus

4.1 Selection du corpus analysé

Nous avons mené une collecte de données dans la presse écrite. À la différence du travail mené par Polguère, nous avons visé le code écrit. Pour ce faire, nous avons sélectionné plusieurs journaux dont l'espagnol est essentiellement l'espagnol général et nous n'avons pas retenu, par exemple, des régionalismes. Ces journaux sont *El País*, *La Vanguardia* , *La Gaceta de Salamanca* et *El Adelanto de Salamanca*.

4.2 Critères suivis

Tous les exemples du corpus ont été, en premier, vérifiés dans des dictionnaires spécialisés de cooccurrences tels que *REDES*, ou bien dans des dictionnaires de l'espagnol usuel, tels que celui de María Moliner et celui de Seco. Deuxièmement, les exemples ont été vérifiés dans le principal corpus de référence de l'espagnol contemporain : *CREA* (Corpus de referencia del español actual) de l'Académie Espagnole de la langue. Et finalement, nous avons utilisé des informateurs à l'université.

5 Mode d'analyse des données

Pour la description des données concernant les greffes collocationnelles de notre corpus, nous avons suivi les conventions d'écriture et le mode d'analyse suivants :

Répérage et extraction des greffes collocationnelles; chaque énoncé étant numéroté de la façon suivante : *g-1*, *g-2*, *g-n*… La greffe est en gras dans l'énoncé.

→ <énoncé qui était la cible vraisemblable du locuteur>
<lien de fonction lexicale qui doit être activé pour produire cet énoncé>
↑ <énoncé qui est la source de la greffe>
<lien de fonction lexicale qui doit être activé pour produire cet énoncé>
Commentaire. <commentaire de l'analyse proposée>

Ex :

(g-0) *Nadal aplasta a Federer. Nadal* ***infligió una lección*** *al número uno mundial* (*La Vanguardia*, 10-06-08)

→ ⌜*dar una lección*⌝
Oper$_1$(*lección*) = *dar* [ART ~ *à* N]
↑ *infligir una derrota <castigo>*
Oper$_1$(*derrota*) = *infligir* [ART ~ *à* N]
Commentaire. Nous considérons que le locuteur a visé la locution ⌜*dar una lección*⌝ dont le sens est ⟨dar un castigo ejemplar⟩. La source de cette greffe est la collocation *infligir una derrota*, *un castigo*, par conséquent, la composante sémantique ⟨castigo⟩ est partagée aussi bien par la locution que par la collocation.

(g-1) ***Segar la libertad de expresión*** *es otra forma de matar* (*La Gaceta de Salamanca*, 09-06-08)

→ *privar de libertad, limitar/negar/coartar la libertad*
LiquFunc$_1$(*libertad*) = *privar* [de ~]
↑ *segar la vida, <la esperanza>, <la ilusión>*
LiquFunc$_1$(*vida*) = *segar* [ART ~ *à* N]
Commentaire. En espagnol, le vocable VIDA regroupe plusieurs lexèmes VIDA dont l'un a le sens de ⟨cualquier cosa que se considera esencial o muy importante para la vida o el desarrollo de una persona o de una colectividad⟩. Par conséquent, il y a un lien sémantique entre les bases de la greffe source et la greffe cible.

6 Typologie et analyse des greffes collocationnelles

6.1 Greffes évidentes

(g-2) *A mí que no tenga tan* ***robusta la fe*** *como me gustaría, pensar que hay dios me tranquiliza en los peores momentos*. (Marta Robles, journaliste, *La Gaceta de Salamanca*, 13-01-09)

→ *fe profunda/sólida/inquebrantable*
Magn(*fe*) = *profunda, sólida <inquebrantable>*
↑ *espíritu <moral > robustecido*
Magn(*espíritu*) = *robustecido, fuerte <inquebrantable>*
Commentaire. Il y a un lien sémantique évident entre la base de la collocation cible et la base de la source. Elles appartiennent au même champ sémantique.

(g-3) *Ana Ivanovic* ***asume el trono*** *de Justine Henin*

→*ocupar el trono*
IncepOper$_1$(*trono*) = *ocupar* [ART ~]
↑ *asumir la responsabilidad <el compromiso> <el peso> <la carga>*
Real (*responsabilidad*) = *asumir* [ART ~]
Commentaire. Le lexème TRONO connote une idée de responsabilité et de charge, il reste donc sémantiquement proche du sens de la base de la collocation qui est la source de la greffe.

(g-4) *El rey se negó a* ***interceptar el nuevo estatuto****. A don Juan Carlos le pidieron personas notables del Reino que* ***interceptase la aprobación*** *del nuevo Estatut de Catalunya*. (*La Vanguardia*, 02-11-08)

→*obstaculizar/impedir la aprobación*
LiquFunc$_1$(*aprobación*) = *obstaculizar* [ART ~]*, impedir* [ART ~]
↑ *source pas évidente*
Commentaire. Nous pensons, comme nous l'avons déjà signalé ci-dessus, que la source de cette greffe collocationnelle n'est pas évidente. Cependant, une interprétation serait possible. On pourrait donner comme source une collocation qui prendrait comme base l'ASém Y de *aprobación*, dans ce cas *estatuto*, ayant comme composante sémantique (texte écrit) et que l'on pourrait alors relier sémantiquement à *comunicación*, *mensaje* qui, eux, constituent la base d'une collocation à laquelle l'application de la FL **LiquFunc$_1$** donnerait comme *valeur interceptar*.

LiquFunc$_1$(*comunicación*) = *interceptar* [ART ~]
LiquFunc$_1$(*mensaje*) = *interceptar* [ART ~]

(g-5) *La principal (medida) que se les autorice a* ***incurrir en déficit*** *mientras dure la crisis*. (*El País*, 29-11-08).

→ *contraer déficit*
IncepOper$_1$(*déficit*) = *contraer*
↑ *incurrir en una falta <un delito>, <un error>*
Oper$_1$(*falta*) = *cometer* [ART ~] *incurrir en* [ART ~]*, caer en* [ART ~]
Commentaire. *Déficit* implique (falta), (carencia), (deficiencia), (insuficiencia), (no completo), et a une connotation négative. *Error* et *falta* impliquent aussi un manque. Il y a donc, encore une fois, un lien sémantique entre les bases.

6.2 Collocations perçues dans un premier temps comme étant des greffes, mais qui semblent très largement utilisées

(g-6) La temporada turística está siendo muy buena, y *se ha movilizado una* **importante *bolsa de visitantes* del norte** (*La Gaceta de Salamanca*, 08-06-08)

→ *grupo de visitantes*
Mult(*visitantes*) = *grupo* [*de* ~]
↑ *bolsa de personas*
Mult(*personas*) = *bolsa* [*de* ~]

→ *nutrido grupo de visitantes*
Magn(*grupo*) = *numeroso, amplio, nutrido*
↑ ***Magn*** *générique*

Commentaire. Dans l'expression mise en gras, il y a deux greffes collocationnelles. Dans la première, le locuteur a mal choisi le collocatif de *visiteurs* exprimant le sens ⟨ensemble régulier de…⟩ correspondant à la FL **Mult**. L'origine de la greffe est le sens très proche des deux bases, car elles sont des quasi-synonymes.
La deuxième greffe collocationnelle concerne la FL **Magn**. C'est une greffe collocationnelle très fréquente en espagnol, aussi bien dans la langue écrite que dans la langue orale à tel point qu'elle pourrait être considérée comme faisant partie de la norme.

(g-7) O su hija Ana, que estaba **completamente emocionada** al escuchar a su madre (*La Gaceta de Salamanca*, 12-06-08)

→ *profundamente emocionada*
Magn(*emocionado*) = *profundamente, sumamente, enormemente, intensamente*
↑ ***Magn*** *générique*
Commentaire. La greffe collocationnelle concerne la FL **Magn**. C'est une greffe collocationnelle très fréquente en espagnol, aussi bien dans la langue écrite que dans la langue orale à tel point qu'elle pourrait être considérée également comme faisant partie de la norme.

6.3 Greffes inverses[1]

(g-8) *La **ofrenda** de Zapatero a su tierra **culmina** una semana que para Salamanca quedará marcada por el cabreo y la envidia*. (*La Gaceta de Salamanca*, 08-06-08)

Commentaire. Le locuteur a utilisé la base *ofrenda* au lieu de la base cible *oferta*. Les deux bases sont sémantiquement très proches car toutes les deux ont comme composante sémantique centrale ⟨presente⟩. La différence est que *ofrenda* a la composante sémantique ⟨devoción⟩ ce qui fait qu'elle soit rattachée au champ sémantique ⟨religion⟩, tandis que *oferta* n'a pas cette composante sémantique. En même temps, nous considérons que la proximité phonologique de ces deux bases a joué un rôle important dans la production de cette greffe.

(g-9) *El Arsenal piensa ceder a Barazite y la Real Sociedad **pide el préstamo** (…) y añadía tener constancia de que la Real ya ha remitido un fax al Arsenal **solicitando la cesión**.* (*La Gaceta de Salamanca*, 29-07-08)

Commentaire. Dans cet exemple, encore une fois, c'est la base qui a été mal sélectionnée par le locuteur. Ce sont deux bases sémantiquement proches. La différence étant que *préstamo* est toujours ⟨préstamo de dinero o de alguna cosa⟩ tandis que *cesión* est ⟨acción de ceder un bien o una persona⟩, por ejemplo en el campo sémantico ⟨deporte⟩.

6.4 Interférences syntaxiques entre deux collocatifs quasi-synonymes

(g-10) *En estos tres años, lo que sí ha aprendido el delantero es a **rehuir de la polémica** (La Vanguardia*, 26-09-08)

→ *rehuir la polémica*

[1] Dans ce travail nous ne nous occupons pas des greffes collocationnelles inverses, que nous considérons comme un cas marqué. Dans notre corpus ce type de greffes constitue un cas marginal.

↑ *huir de la tentación*
Commentaire. Nous croyons que cette greffe collocationnelle est motivée parce que ce sont, d'une part, des collocatifs quasi-synonymes et, d'autre part, par la ressemblance phonologique. *Rehuir* a pour glose sémantique ⟨procurar no encontrarse en cierta situación⟩, et *huir* a pour glose sémantique ⟨hacer por no encontrarse en cierta situación⟩. Il s'agit ici d'une greffe de la combinatoire d'un collocatif sur l'autre.

(g-11) *El plan europeo* ***despierte confianza*** (*La Vanguardia*, 04-10-08)

→ *despierte la confianza de…*
CausFunc$_1$(*confianza*) = *despertar* [ART ~]
↑ *despertar esperanzas*
CausFunc$_1$(*esperanzas*) = *despertar*
Commentaire. La combinatoire de la greffe source a parasité la combinatoire de la greffe cible qui, elle, exige l'article devant *confianza*. Les deux bases sont sémantiquement très proches étant donné que *confianza* est ⟨tener la esperanza de que algo va a venir⟩.

6.5 Similarité phonologique

(g-12) **Corridas dos horas y veinte minutos** de partido

→ *transcurrir/pasar dos horas y veinte minutos*
Func$_0$(*hora*) = *transcurrir, pasar*
↑ *correr/pasar/transcurrir el tiempo*
Func$_0$(*tiempo*) = *correr, pasar, transcurrir*
Commentaire. Nous considérons que dans cet exemple c'est surtout la similarité phonologique qui a joué entre les deux collocatifs. Outre la similarité phonologique, les deux bases sont, encore une fois, très proches sémantiquement.

7 Conclusions

Dans ce travail, l'analyse des exemples de notre corpus nous permet de tirer les conclusions suivantes :

1. La greffe type concerne le collocatif et non la base de la collocation. Les greffes inverses sont un cas marginal.
2. La base de la cible et la base de la source sont sémantiquement très proches.
3. La même fonction lexicale est impliquée dans la cible et dans la source de la greffe.

Ces conclusions permettent, à leur tour, de confirmer les prédictions faites par Alain Polguère sur les greffes collocationnelles dans des langues autre que le français, ainsi que de corroborer la validité des notions théoriques et du mode d'analyse.

Nous avons aussi dans notre corpus des greffes portant sur des locutions. Il faudra, par la suite, vérifier si le mode d'analyse suivi pour la description des greffes collocationnelles est valable aussi pour les locutions. Le fait de compléter l'analyse des greffes collocationnelles en ajoutant l'analyse des locutions permettra, à notre avis, de mieux cerner le statut linguistique de ce phénomène.

Références bibliographiques

Alonso Ramos, Margarita. 2004. *Las construcciones con verbo de apoyo*, Visor Libros, Madrid.

Bosque, Ignacio. 2004. *Diccionario combinatorio del español contemporáneo* (REDES), SM, Madrid.

Mel'čuk, Igor A., Arbatchewsky-Jumarie, Nadia, Elnitsky, Léo, Iordanskaja, Lidjia & Lessard, Adèle. 1984. *Dictionnaire explicatif et combinatoire du français contemporain: Recherches lexico-sémantiques* volume 1, Presses de l'Université de Montréal, Montréal.

Mel'čuk, Igor A., Arbatchewsky-Jumarie, Nadia, Dagenais, Louise, Elnitsky, Léo, Iordanskaja, Lidjia, Lefebvre, Marie Noelle & Mantha, Suzanne. 1988. *Dictionnaire explicatif et combinatoire du français contemporain: Recherches lexico-sémantiques* volume 2, Presses de l'Université de Montréal, Montréal.

Mel'čuk, Igor A., Arbatchewsky-Jumarie, Nadia, Iordanskaja, Lidjia & Mantha, Suzanne. 1992. *Dictionnaire explicatif et combinatoire du français contemporain: Recherches lexico-sémantiques* volume 3, Presses de l'Université de Montréal, Montréal.

Mel'čuk, Igor A., Arbatchewsky-Jumarie, Nadia, Iordanskaja, Lidjia, Mantha, Suzanne & Polguère, Alain. 1999. *Dictionnaire explicatif et combinatoire du français contemporain: Recherches lexico-sémantiques* volume 4, Presses de l'Université de Montréal, Montréal.

Mel'čuk, Igor A. & Polguère, Alain. 2006. Dérivations sémantiques et collocations dans le DiCo/LAF. In Peter Blumenthal & Franz Josef Hausmann (ed.), *Collocations, corpus, dictionnaires*, vol. 150, *Langue Française*, 66-83. Larousse, Paris.

Mel'čuk, Igor A., Clas, André, & Polguère, Alain. 1995. *Introduction à la lexicologie explicative et combinatoire*, Duculot, Louvain-la-Neuve.

Mel'čuk, Igor A., & Polguère, Alain. 2007 *Lexique actif du français*, De Boeck & Larcier, Bruxelles.

Moliner, María. 1998. *Diccionario del uso del español* (DUE), Gredos, Madrid.

Polguère, Alain. 2008. *Lexicologie et sémantique lexicale. Notions fondamentales*, nouvelle édition revue et augmentée, coll. « Paramètres », Les Presses de l'Université de Montréal, Montréal.

Polguère, Alain. 2007. *Soleil insoutenable* et *chaleur de plomb* : le statut linguistique des greffes collocationnelles. Version officielle non publiée de l'article. Téléchargeable sur www.olst.umontreal.ca/

Polguère, Alain. 2003. Collocations et fonctions lexicales : pour un modèle d'apprentissage. In Francis Grossmann & Agnès Tutin (ed.), *Les Collocations. Analyse et traitement*, coll. "Travaux et Recherches en Linguistique Appliquée", E:1, 117-133. De Werelt, Amsterdam.

Seco, Manuel, Andrés, Olimpia & Ramos, Gabino. 1999. *Diccionario del español actual*, Aguilar, Madrid.

On the place of information structure in a grammar

Robert D. Van Valin, Jr.
Heinrich-Heine-Universität Düsseldorf
University at Buffalo, The State University of New York
Max Planck Institute for Psycholinguistics, Nijmegen

The question, 'where does information structure go in the structure of a grammar' is first and foremost an architectural question, and different grammar architectures give different answers. In this talk I will give one possible answer, based on Role and Reference Grammar [RRG] (Van Valin 2005), a monostratal (non-derivational) linking theory. In RRG there is a direct linking between the semantic representation of a sentence and its syntactic representation, and information structure plays a role in this linking. The steps in the linking algorithm mapping semantics into syntax will be specified, and it will be shown that that information structure notions can play a role at every step. Furthermore, aspects of the interaction between discourse context and information structure will be captured using a version of Discourse Representation Theory.

Géométriser le sens lexical

La synonymie comme accès à la sémantique

Fabienne Venant
Loria / Université Nancy2
615, rue du Jardin Botanique
54600 Villers lès Nancy Cedx
fabienne.venant@loria.fr

Abstract

Cet article présente une modélisation du sens lexical, centrée sur le phénomène de la polysémie. Ce modèle utilise la relation de synonymie comme accès aux informations lexico-sémantiques, et propose une réponse géométrique à la question de la représentation du sens. L'étude approfondie des relations de synonymie permet de mettre en évidence à la fois le fonctionnement des unités polysémiques prises individuellement, et leur place dans l'organisation globale du lexique. Il s'agit donc de construire des espaces sémantiques à différentes échelles, pouvant rendre compte de la sémantique d'une unité donnée ou permettre un accès à la structuration sémantique du lexique. La ressource ainsi construite permet d'accéder de façon synthétique à la sémantique des entrées d'un dictionnaire de synonymes. Des pistes sont envisagées vers une exploration plus globale du lexique ainsi que vers une articulation des niveaux conceptuels et sémantiques.

1 Introduction

Le travail présenté ici s'inscrit dans un cadre plus général de modélisation des processus de construction du sens au sein d'un énoncé (Victorri & Fuchs, 1996, Venant, 2006). Ces travaux nous ont amenés à accorder une place centrale aux phénomènes de la polysémie et de la synonymie. Ces phénomènes touchent une grande diversité d'unités linguistiques. Tenter de les formaliser suppose d'une part de définir le sens lexical, et la représentation que l'on veut en donner, et d'autre part de rendre compte des relations sémantiques entre les unités lexicales. Nous avons tenté de donner une réponse géométrique à ces questions, à travers la relation de synonymie. Nous associons donc à chaque vocable polysémique un espace sémantique continu, dans lequel les sens des différentes acceptions s'organisent selon leurs proximités sémantiques. Ces espaces sont construits automatiquement, pour n'importe quelle entrée du Dictionnaire Electronique des Synonymes (DES[2]). Le DES est pour cela modélisé sous forme de graphe. L'algorithme de construction des espaces sémantiques repose sur l'analyse de ce graphe.

2 Sens lexical

Le modèle présenté ici un modèle continu du sens lexical. Notons que quand nous parlons de continu, il s'agit bien de caractériser le modèle du sens utilisé et non le sens lui-même. Comme le remarque Polguère (2002): « C'est un lieu commun de dire qu'il n'existe pas de consensus sur ce qu'il faut entendre par sens linguistique. Toute définition de cette notion ne peut qu'être partielle en regard des différentes façons d'aborder l'étude du contenu des énoncés. » Nous travaillons ici dans le cadre de la construction dynamique du sens (Victorri & Fuchs, 1996) qui considère que le sens d'une unité linguistique dans un énoncé

[2] http://www.crisco.unicaen.fr/cgi-bin/cherches.cgi.

donné est le résultat de l'interaction entre un apport sémantique constant associé à cette unité, que l'on peut appeler son *noyau de sens*, ou encore sa *forme schématique*, et le contexte d'énonciation de cette unité. Le contexte d'énonciation comporte à la fois les autres unités linguistiques présentes dans l'énoncé (son cotexte) et la situation extralinguistique dans laquelle cet énoncé est proféré. Il ne s'agit pas de nier le fait que l'expression linguistique fournisse une composante du matériel de base requis pour l'élaboration du sens, mais de souligner le fait qu'elle n'est qu'une parmi ces composantes. Cette position peut-être rapprochée de celle de Croft et Cruse (2004). Pour ces auteurs, les sens sont des choses qu'on élabore en utilisant les propriétés des éléments linguistiques comme indices partiels tout au long de connaissances non linguistiques, mais aussi les informations disponibles à partir du contexte et les connaissances ou conjectures quant à l'état d'esprit du locuteur.

Le noyau de sens n'est donc pas un sens à proprement parler, mais plutôt un schéma de base à partir duquel se construisent les différents sens d'une unité. Pour l'appréhender, il faut tenter de cerner les propriétés du mot lui-même, qui expliquent à la fois qu'il puisse prendre des sens différents selon les énoncés, et en même temps qu'il évoque à lui tout seul la classe d'objets à laquelle on l'associe spontanément. Pour en donner une idée plus intuitive, disons que le noyau de sens est précisément ce qu'on cherche à appréhender lorsqu'en consultant un dictionnaire, on parcourt l'ensemble de l'article concernant un mot pour se faire une idée de la façon dont il « fonctionne » en contexte.

Prenons par exemple le nom *cours*. L'examen des différentes définitions constituant l'article associé à ce nom dans le Trésor de la Langue Française Informatisé (TLFI)[3], nous permet de dégager le noyau de sens suivant :

Développement (écoulement) continu entre des limites précises,
qui peut ensuite se décliner selon trois perspectives :

- une perspective spatiale : *cours d'une rivière, le cours du sang dans notre corps, cours du soleil, cours de Vincennes ;*
- une perspective temporelle : (temps lui-même*) cours des saisons*, (suite de choses ou d'évènements), *cours des pensées*, (valeurs) *cours du coton ;*
- une perspective que le TLFI qualifie de spatio-temporelle, mais qu'on pourrait plutôt qualifier de notionnelle : *cours de latin*, *donner, faire, suivre un cours*, par métonymie, traité, manuel, établissement - (en parlant d'autres activités) *au long cours.*

En contexte, ce noyau de sens entre en interaction avec les autres éléments linguistiques pour donner naissance à un sens précis. Les énoncés 1 et 2 illustrent ce phénomène. Dans l'énoncé 1, l'interaction entre les noyaux de sens de *cours*, de *rivière* et de *suivre* conduit à la sélection de la perspective spatiale. L'énoncé 2, quant à lui, met en jeu la perspective spatio-temporelle de *cours*.

(1) *La rivière suit son cours.*
(2) *L'étudiant suit un cours.*

3 Des espaces sémantiques pour représenter le sens lexical

Il s'agit ici d'associer à chaque vocable polysémique un espace continu qui rende compte de son noyau de sens. On veut capturer et représenter son organisation sémantique : les différentes nuances de sens qu'il peut prendre, leurs proximités sémantiques, comment on peut passer continûment de l'une à l'autre. L'idée est d'utiliser la relation de synonymie comme accès à cette structure sémantique.

Les espaces sémantiques sont construits automatiquement à partir d'un graphe de synonymie tiré du DES. La méthode utilisée est celle initialement proposée par Ploux et Victorri (1998). On explore ce dictionnaire sous la forme d'un graphe : les sommets sont les expressions linguistiques. Un lien est tracé entre deux expressions lorsque le dictionnaire signale qu'elles sont synonymes. L'idée sous-jacente à la construction des espaces sémantiques est qu'un synonyme n'est généralement pas suffisant pour définir un sens précis d'une expression. Ainsi, au sein de sa synonymie avec *cours*, *fil* est à la fois synonyme de

[3] http://atilf.atilf.fr/tlf.htm.

courant et de *déroulement*, ce qui correspond à deux sens distincts de *cours*, l'un lié à la perspective spatiale (ou notionnelle), l'autre à la perspective temporelle. Or les points de l'espace sémantique doivent correspondre à des sens précis de l'unité. C'est pourquoi nous avons recours à la notion de clique. Une clique est un sous-graphe complet maximal, c'est-à-dire un ensemble de sommets, le plus grand possible, reliés deux à deux. Voici par exemple la liste des cliques qui, dans le sous-graphe formé par *cours* et tous ses synonymes, contiennent le nom *fil*. Chaque clique correspond à une nuance possible de sens pour *cours* : < *courant ; cours ; fil ; flot* >, < *courant ; cours ; fil ; succession* > et < *cours ; déroulement ; enchaînement ; fil ; succession ; suite* >

L'espace sémantique est une projection en deux dimensions du nuage formé par les cliques dans l'espace multidimensionnel engendré par les synonymes de l'exspression considérée. Il est muni de la métrique du chi2. C'est elle qui s'est en effet avérée la plus efficace pour obtenir une représentation respectant la notion intuitive de proximité entre sens. La Figure 1 présente l'espace sémantique de *cours*.

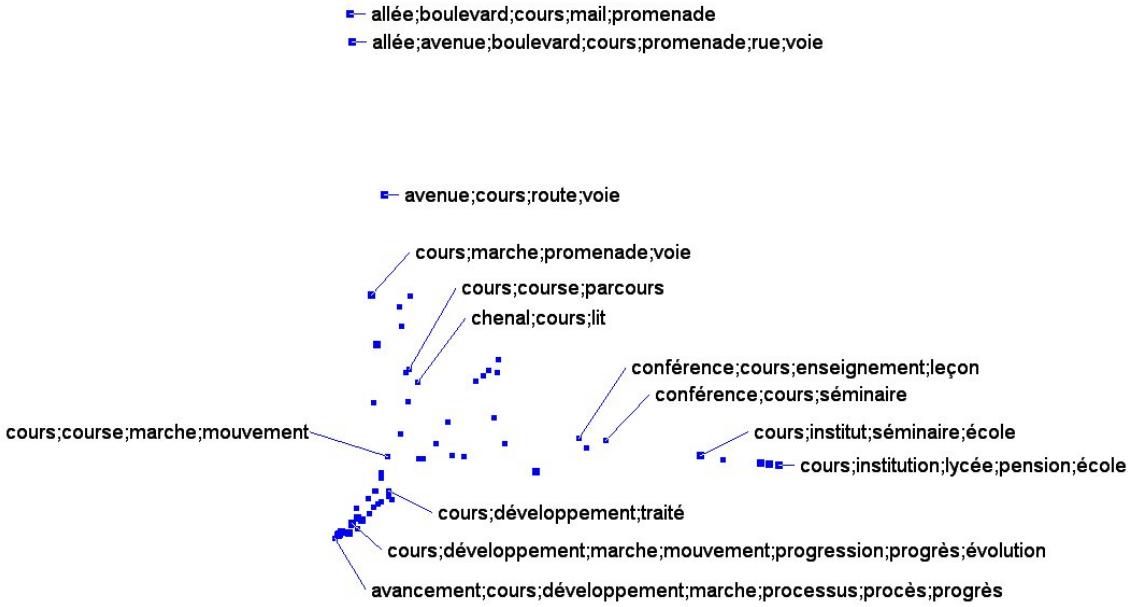


Figure 1. Espace sémantique associé à *cours*

4 Continuum de sens

Les espaces sémantiques ainsi construits permettent de rendre compte d'un continuum entre les différents sens d'une unité, et par là même d'appréhender le noyau de sens de cette unité. On voit par exemple, sur la Figure 1, que l'espace sémantique de *cours* fait clairement apparaître deux branches, l'une (partie haute de l'espace sémantique) correspondant à la perspective spatiale, avec des cliques comme < *allée ; boulevard ; cours ; mail* > ou < *chenal ; cours ; voie* >, l'autre (vers la droite) correspondant à la perspective notionnelle, avec des cliques comme < *conférence ; cours ; séminaire* > ou < *cours ; mouvement ; école* >. On passe continûment d'une branche à l'autre, via la zone centrale, qui rassemble des cliques relevant plutôt de la perspective temporelle comme < *cours ; durée ; courant* >; ou < *cours ; fil ; succession* >. Une telle représentation permet, entre autres, d'expliciter la relation de famille unissant deux sens d'une unité donnée. Ainsi on peut ainsi exhiber une suite de sens intermédiaires permettant de passer continûment du sens de *cours* dans « la rivière suit son cours » à celui qu'il prend dans un étudiant «L'étudiant suit un cours », via les cliques suivantes :

< *chenal ; lit ; cours* >
< *chenal ; courant ; cours* >
< *cours ; course ; marche ; mouvement* >
< *courant ; cours ; flot ; mouvement* >

< *courant ; cours ; fil ; succession* >
< *courant ; cours ; mouvement* >
< *cours ; mouvement ; école* >
< *cours ; classe ; leçon ; école* >
< *cours ; enseignement ; leçon ; conférence* >

5 Caractériser la polysémie

Partant du fait que le sens d'une expression linguistique dans une occurrence donnée dépend en partie de ce qu'apporte cette expression elle-même de constant, quel que soit le contexte, et en partie de ce qui est fonction du contexte, on peut classer les expressions suivant l'importance relative de ces deux facteurs. A un extrême, le contexte ne joue aucun rôle : l'expression est monosémique ; son sens est le même dans tous les énoncés, ce sens étant donc entièrement défini par l'apport propre de l'expression (exemples : *tournevis*, *hectolitre*…). A l'autre extrême, on trouve les homonymes « purs », dont l'apport constant, commun à tous les emplois, est effectivement nul, puisque le sens peut changer radicalement suivant les énoncés (exemples : *avocat*, v*ol*…). Entre ces deux extrêmes, se trouve le cas général de la polysémie, avec des cas qui tendent vers la monosémie, quand le contexte ne joue qu'un rôle minime (tous les sens recensables sont très proches les uns des autres), et d'autres vers l'homonymie, quand l'apport propre constant est très faible[4].

Prenons quelques exemples pour illustrer ce dernier point. Soit le mot *bureau*. Il possède quatre sens principaux : un meuble (ex. : *s'asseoir à son bureau*), une pièce (ex. : *ouvrir la fenêtre de son bureau*), un établissement (ex. : *le bureau de poste*, *le bureau de tabac*, etc.), une institution (ex. : *le bureau de l'Assemblée*, *le bureau de l'association*, etc.). Ces différents sens sont indéniablement reliés, ce qui signifie que l'on a bien affaire à de la polysémie et non à de l'homonymie. Cependant, quand on essaie de déterminer l'apport propre du mot *bureau* qui est commun à tous ses emplois, on s'aperçoit qu'il est très ténu : une vague notion d'activité d'écriture, qu'il semble difficile de rendre opératoire dans un calcul effectif du sens de *bureau* en contexte.

Prenons maintenant le mot *livre*. Ce nom est particulièrement intéressant car il relève d'une polysémie tout à fait particulière, que Kleiber (1999) appelle la polysémie logique, et qui se situe à la frontière entre polysémie lexicale et variation contextuelle. Le problème classique posé par *livre* est qu'il semble présenter deux sens principaux, attestés par les dictionnaires, mais que ces deux sens possèdent des propriétés sémantiques qui empêchent de les considérer comme les différents sens d'un polysème standard. Le TLFI distingue ainsi deux sens pour *livre. L*'un désigne le livre en tant qu'objet, assemblage de feuilles destiné à être lus. Dans ce sens *livre* désigne un objet matériel. L'autre désigne le livre en tant que contenu, l'œuvre en elle-même. Dans ce sens, *livre* désigne un objet abstrait, informatif. On peut voir en *livre* un polysème standard, prenant selon le contexte un sens ou l'autre. Mais la particularité de *livre* (et d'autres noms comme *poulet, film, banque* ou les noms de ville) est, d'une part, que ses différents sens peuvent se retrouver unifiés dans un sens global, comme dans l'énoncé 3, et que, d'autre part, ils ne sont pas concurrents et peuvent coopérer au sein d'un même énoncé (4 et 5) – ce qui est quasi impossible pour les sens d'un polysème standard.

(3) *J'ai acheté un nouveau livre.*
(4) *Le petit livre tout jauni que tu m'as prêté est particulièrement intéressant.*
(5) *C'est un livre lourd à trimbaler mais très facile à lire.*

On a donc affaire dans ce cas à une polysémie beaucoup plus proche de la monosémie que de l'homonymie. Certains auteurs (Croft & Cruse 2004, Jayez 2008) parlent alors de facettes.

Un des intérêts des espaces sémantiques que nous construisons est de pouvoir rendre compte de ce continuum entre polysémie et homonymie. Cela permet de pourvoir traiter les cas intermédiaires, comme *livre* et *bureau*, sans avoir à décider à l'avance d'un éventuel typage particulier. Nous ne sommes bien sûr pas en mesure de représenter correctement les monosèmes, car ils correspondent à des sens très spécifiques et ne possèdent donc que très peu de synonymes. Notre algorithme ne permet pas de leur associer

[4] Notre approche de la polysémie est détaillée dans Jacquet et al, 2005.

un espace sémantique. Certains (comme t*ournevis*) sont même absents du DES. Nous obtenons en revanche des résultats intéressants quand il s'agit de caractériser différents types de polysèmes. Nous présentons à titre d'exemple les espaces sémantiques associés aux noms *vol*, *bureau* et *livre*. Les espaces sémantiques rendent clairement compte des distinctions que nous venons de mentionner :

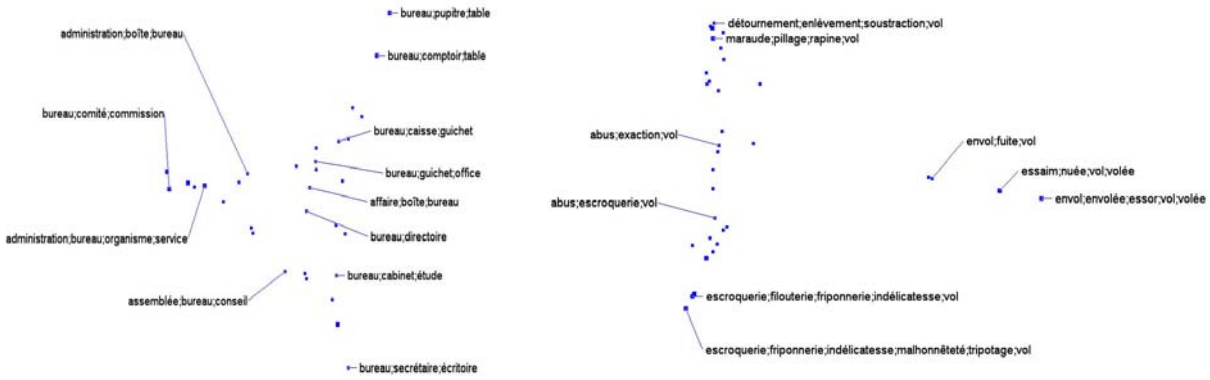


Figure 2. Espaces sémantiques associés à *bureau* et *vol*

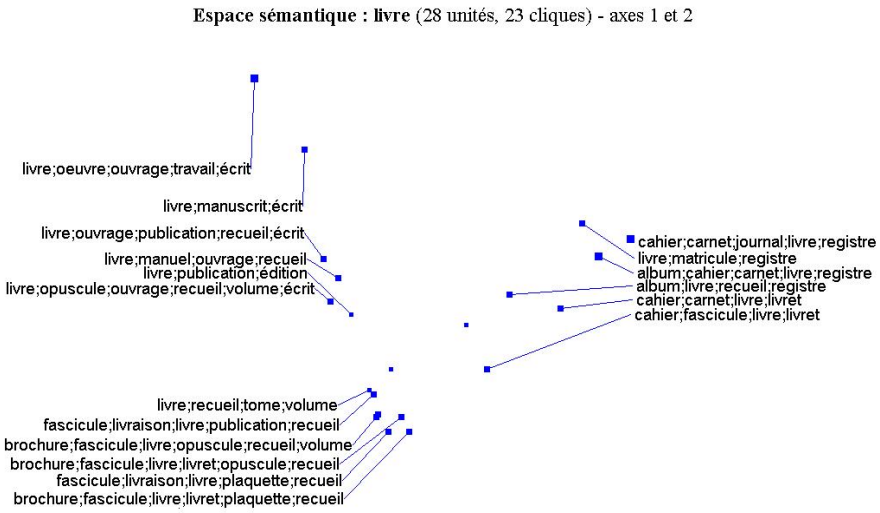


Figure 3. Espace sémantique associé à *livre*

- Figure 2 : L'espace sémantique de *bureau* est clairement continu. Même s'il est difficile à déterminer, le noyau de sens de *bureau* existe et établit des liens sémantiques indéniables entre les différents sens de *bureau.* L'espace sémantique de *vol* est, quant à lui, séparé en deux sous-espaces disjoints, sans chevauchement. On repère ainsi deux lexies bien différentes. On est clairement dans un cas d'homonymie.
- Figure 3 : L'espace sémantique de *livre*[5] fait apparaître deux branches. La partie supérieure gauche de l'espace sémantique rassemble les cliques contenant les synonymes comme *oeuvre, travail, écrit, ouvrage*, c'est-à-dire les sens correspondant au contenu d'un livre. La partie droite de l'espace sémantique rassemble les cliques contenant les synonymes comme *album*, *cahier*, *registre,* c'est-à-dire les sens correspondant à des emplois de *livre* comme objet matériel, assemblage de feuille et support d'écriture. Ces deux branches se rejoignent au niveau de la partie centrale inférieure de l'espace sémantique, qui constitue une région de sens intermédiaire entre les

[5] Pour une étude approfondie, voir Venant (2008).

deux précédentes. Les cliques de cette région de l'espace contiennent des procès de type /publication/, comme *livraison*, *plaquette*, *brochure*…. L'accent est mis ici sur l'objet que l'on obtient, le résultat de la publication. La topologie obtenue pour *livre* fait donc état d'un continuum de sens, qui permet de passer progressivement des sens /contenu à l'état pur/, à l'extrémité de la branche gauche, aux sens /objet à l'état pur/, à l'extrémité de la branche droite. La continuité est assurée par les sens /résultat d'une publication/, ce qui paraît tout à fait cohérent puisque l'un des objets de la publication, c'est précisément de faire le lien entre le contenu et la forme. Il est difficile de séparer le contenu et la publication, tout simplement parce que le contenu d'un livre est conçu par l'écrivain comme devant être publié, et il n'est accessible au lecteur que parce qu'il a été publié. De même il est difficile de séparer la publication de l'objet matériel puisqu'il en est le résultat.

On voit sur ces quelques exemples que l'algorithme utilisé permet d'obtenir, de façon totalement automatique, et très rapidement, des représentations sémantiques fiables, fidèles à la nature sémantique de l'unité étudiée. Elles rendent compte des distinctions de sens que l'on peut trouver dans les dictionnaires, tout en proposant un éclairage sémantique différent. Ainsi, dans le cas de *livre*, on voit apparaître une zone de sens centrale, le livre comme résultat d'une publication, dont la spécificité est d'être un sens à la fois intermédiaire et synthétique entre les facettes classiques *objet* et *contenu*. Elle exprime une propriété fondamentale du livre, celle qui distingue le livre d'une simple lettre ou d'un journal intime, à savoir la ***dimension sociale*** du livre. Car le livre est plus qu'un contenu s'appuyant sur une certaine forme. Le livre est avant tout un objet social, doté d'un contenu destiné à être rendu public. C'est en intégrant cette dimension, apparemment oubliée des analyses classiques, que l'on pourra interpréter des expressions comme « un livre à succès », « un livre qui tranche, qui innove », « un livre qui est passé inaperçu. »

6 Zones de sens dans un espace sémantique

Les espaces sémantiques tels que nous les construisons permettent une étude approfondie de la sémantique d'une unité donnée, en mettant au jour une structure sémantique sous-jacente. Ils sont tels quels très utilisés par les internautes s'interrogeant sur les différents acceptions d'une unité et/ou les synonymes pertinents dans un contexte donné. Ils sont aussi utilisés par les linguistes s'intéressant aux caractéristiques sémantiques des unités lexicales. François et Senechal (2004) les utilisent par exemple pour caractériser les différents foyers de polysémie d'un verbe. Nous avons pour notre part cherché à les exploiter pour modéliser le sens pris par un vocable polysémique dans un énoncé (Venant 2004, 2006, 2008, Jacquet, 2003). L'idée est de modéliser l'interaction entre le noyau de sens et le cotexte, afin de déterminer la zone de l'espace sémantique qui correspond au sens pris par l'unité considérée dans le cotexte étudié. Nous ne détaillons pas ici le calcul effectué. Disons simplement que nous calculons un taux d'affinité entre le cotexte et chaque clique de l'espace sémantique. Ce calcul est effectué à partir de données de cooccurrences issues d'un corpus. Les taux d'affinité calculés permettent d'obtenir une « déformation » de l'espace sémantique. La zone de déformation correspond au sens pris par l'unité étudiée dans le cotexte considéré. Le mode de représentation, continu, que nous avons choisi prend ici tout son intérêt car il nous permet de modéliser le sens par une région de l'espace sémantique. Le fait d'utiliser une région, et non un point, permet de rendre compte de tous les cas de figure interprétatifs. Une région étroite correspond à un sens précis, une région étendue à un sens plus indéterminé, une région non connexe à une ambiguïté.

Ce mode de représentation permet ainsi de rendre compte de la polysémie logique[6]. La figure 4 montre des zones de sens calculées sur l'espace sémantique de *livre*. La partie gauche de la figure 4 présente les zones de sens activées par l'adjectif *ancien* (correspondant donc aux sens possibles pour *livre* dans *un livre ancien* ou *un ancien livre*). On y voit deux régions nettement séparées. On est donc en présence d'une ambiguïté. *Ancien* sélectionne tantôt le sens /contenu/ de *livre* (comme dans « J'aime son dernier roman mais je trouve ses anciens livres mieux écrits »), tantôt le sens /objet/ de *livre* (comme dans « J'ai trouvé un joli livre ancien chez l'antiquaire »). C'est le contexte élargi qui permet de trancher.

[6] voir Venant (2008).

La partie droite de la figure 4 montre les zones de sens obtenues pour l'adjectif *prochain*. On y voit comment l'unification des facettes /contenu/ et /objet/ se fait, via l'activation de la zone centrale de l'espace correspondant au livre en tant que publication. Dans l'expression *prochain livre*, on réfère à la fois à l'objet manufacturé, résultat de la publication, et au contenu, comme en témoigne l'énoncé : « Entre deux livres, et plus précisément dans ce temps où il n'y a encore pas de place bien définie pour cet objet inexistant qu'est « le prochain livre », c'est là que viennent s'accumuler les notations, les intentions[…] » (Genèse du romain contemporain, D.Ferrer et B. Boie, p44)

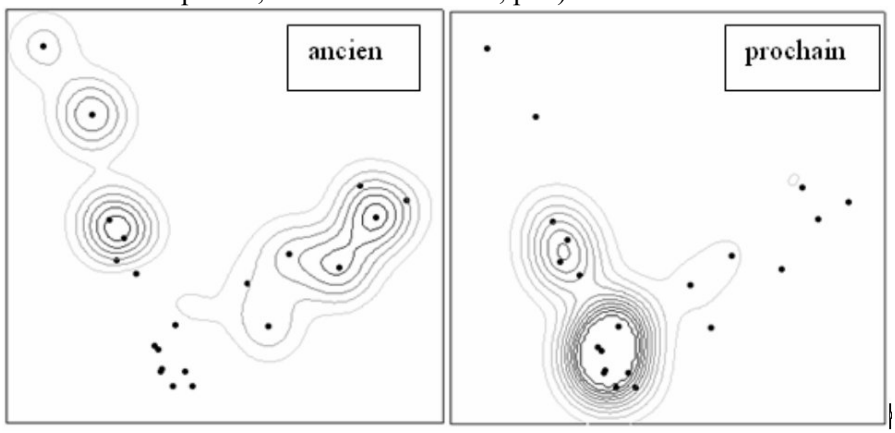


Figure 4. Zone de sens dans l'espace sémantique de *livre*

7 Granularité et changement d'échelle

Notre modèle repose donc sur le choix d'une représentation continue du sens lexical, et en particulier sur le choix des cliques comme unité minimale de représentation sémantique et sur le tuilage continu de l'espace sémantique qu'elles réalisent. Cependant, les cliques représentent des nuances de sens très précises, souvent très proches les unes des autres. Cette subtilité sémantique génère une redondance qui peut poser problème dans certaines applications. Une granularité aussi fine rend difficile, voire impossible, l'étude des relations entre unités lexicales, ainsi que l'articulation entre niveau conceptuel et sémantique. La représentation continue du sens est très adaptée pour certaines tâches, comme l'étude des ambiguïtés ou la recherche d'un synonyme pertinent pour une unité donnée dans un contexte donnée. En revanche, une désambiguïsation à grande échelle, ou un étiquetage sémantique, vont demander l'utilisation d'étiquettes macroscopiques, et le regroupement des cliques en zones de sens correspondant, par exemple, aux définitions d'un dictionnaire. Nous cherchons donc désormais à avancer vers une représentation à granularité variable, où l'on pourrait, selon les besoins, accéder aux nuances de sens représentées par les cliques, ou à des sens macroscopiques correspondant à des regroupements de cliques. Plusieurs pistes sont en cours. Elles visent à repérer des zones denses dans les nuages de points formés par les cliques, soit par des méthodes géométriques de clusterisation, soit par interaction avec d'autres ressources sémantiques.

La possibilité de grouper les cliques permet aussi d'envisager des changements d'échelle. La clique constitue en effet une unité structurelle trop fine pour une étude globale des relations sémantiques au sein du lexique. Le changement de granularité permet de prendre du recul et d'étudier la place des unités lexicales dans l'organisation globale du lexique.

7.1 Un espace global

La représentation que nous venons de décrire est centrée sur une seule unité linguistique, par exemple le nom *cours.* Les synonymes de *cours* n'interviennent qu'en tant qu'ils permettent de distinguer ses différentes acceptions : ils restent cantonnés à l'intérieur de l'espace sémantique de *cours*. Les seuls sens de ces synonymes qui sont pris en considération sont ceux qu'ils partagent avec l'unité sur laquelle on s'est focalisée. Il s'agit donc d'une représentation tronquée de la relation de synonymie, puisque seul l'un des synonymes (en l'occurrence *cours*) se trouve pleinement représenté. Pour représenter plus fidèlement les

relations entre *cours* et ses synonymes, il faut donc passer d'un espace sémantique *local*, associé à une unité, à un espace sémantique *global*, qui replace l'ensemble des unités concernées dans le réseau complet ou dans un sous-réseau paradigmatique. D'une manière générale, ce n'est que dans un espace global que l'on peut obtenir une représentation fidèle des relations de synonymie entre plusieurs unités lexicales.

La clique a montré son efficacité en tant qu'outil d'exploration d'un graphe lexical. Le problème, on l'a dit est celui du niveau de granularité. Si on travaille, par exemple, sur un graphe adjectival[7], on se retrouve avec plus de 11000 cliques. Un si grand nombre de cliques ne permet d'envisager ni l'exploration, ni l'exploitation, d'un tel espace sémantique. L'idée est donc de repérer des branches de cliques correspondant à des sens très proches mais pas forcément deux à deux. Pour les mettre en évidence, nous avons défini un outil géométrique, la **boule**, à partir duquel nous définissons et construisons des **branches** de cliques. On forme autour de chaque clique une boule contenant ses voisines les plus proches, à l'exclusion de celles qui n'ont aucun synonyme commun avec elle. Les branches que nous cherchons à mettre en évidence sont alors des rassemblements de boules, rassemblées en fonction du nombre de synonymes qu'elles ont en commun.

La mise en œuvre de ces outils sur un graphe adjectival (Venant, 2006) nous a permis d'une part de caractériser les grandes classes adjectivales traditionnellement distinguées : qualificatifs, intensifs (Romero 2004), relationnels (Daille 2001), primaires (Borodina 1963, Goes 1999, Noailly 1999), et d'autre part de montrer que, d'un point de vue théorique, il ne fallait pas chercher à classer les adjectifs eux-mêmes, mais leurs emplois, un même adjectif pouvant appartenir à différentes classes suivant ses emplois. Ainsi nous avons pu montrer que la plupart des adjectifs, même les plus qualificatifs, possèdent des emplois intensifs, comme par exemple *méchant* dans des emplois tels que *une méchante voiture* (intensif positif) ou *un méchant costume de laine* (intensif négatif). Il existe d'autre part un continuum entre les différents types d'emplois adjectivaux, qu'on se situe au niveau d'un adjectif donné comme *intime* : de *l'ami intime* (qualificatif) à *l'intime conviction* (intensif), ou encore *adolescent* : de *l'amour adolescent* (qualificatif) à *un groupe adolescent* (relationnel), ou que l'on prenne le lexique dans son ensemble : des adjectifs qualificatifs intenses comme *bouillant* (*eau bouillante*) aux intensifs purs comme *énorme* (*énorme envie*). Ceci explique pourquoi certains linguistes, comme Bartning & Noailly (1993), ont tant peiné en cherchant à établir des frontières nettes, notamment entre adjectifs qualificatifs et adjectifs relationnels. Nos résultats nous amènent plutôt à adhérer à la proposition de Goes (1999) ou de Romero (2004) de parler d'adjectifs « statistiquement relationnels » (ou « statistiquement intensifs ») pour caractériser des adjectifs comme *procédural* (resp. *extrême*), dont les emplois sont majoritairement relationnels (resp. majoritairement intensifs), sans que cela exclue la possibilité de trouver ces adjectifs dans des emplois purement intensifs, ex. : *une lenteur procédurale* (resp. purement qualificatifs, ex. : *une expérience extrême*).

7.2 Vers une représentation conceptuelle

Une autre piste pour représenter les sens macroscopiques d'une unité lexicale est d'apparier les cliques d'une unité donnée aux descriptions sémantiques d'une approche discrète. Nous avons initié le travail en utilisant les définitions du TLFI. Pour cela, nous exploitons le travail réalisé par Falk et al (2009) qui vise à apparier, via un calcul de similarités, les sens possibles d'un lexème avec les ensembles de synonymes appropriés. L'idée est de projeter sur les cliques les appariements synonymes/définitions ainsi obtenus. Falk et al se sont pour l'instant limitées au lexique verbal.

A titre de première expérimentation, nous avons calculé pour chaque clique du verbe *abandonner* un taux d'affinité avec une définition du TLFI. Le calcul est le suivant: soit c une clique et d une définition, on note Sc l'ensemble des synonymes appartenant à la clique c, et Sd l'ensemble des synonymes appariés à la définition d par Falk et al. Le taux d'affinité T(c, d) entre la clique c et la définition d est défini par $T(c,d)= |Sc \cap Sd|/|Sc|$

La Figure 5 montre les projections obtenues automatiquement : dans la partie gauche, on marqué d'un astérisque les cliques ayant un taux d'affinité supérieur à 60% avec la définition «Renoncer à un pouvoir,

[7]Nous avons extrait du DES un graphe adjectival. Il comporte ses 198 549 arcs pour 49133 sommets et contient 11900 cliques.

à des droits, à la possession d'un bien ou à l'utilisation d'une chose ». Dans la partie droite, on a marqué les cliques ayant un taux d'affinité supérieur à 60% avec le sens « Quitter un lieu. »

On voit que l'utilisation du taux d'affinité peut constituer un bon moyen de repérer automatiquement dans le nuage de points formé par les cliques, les sens macroscopiques correspondant aux définitions du TLFI. On peut ainsi espérer structurer de façon automatique les espaces sémantiques, repérer et caractériser automatiquement les différentes lexies d'un polysème et enrichir notre représentation sémantique avec les informations spécifiées dans les définitions du dictionnaire (domaine, constructions, sous-catégorisation, combinatoire…).

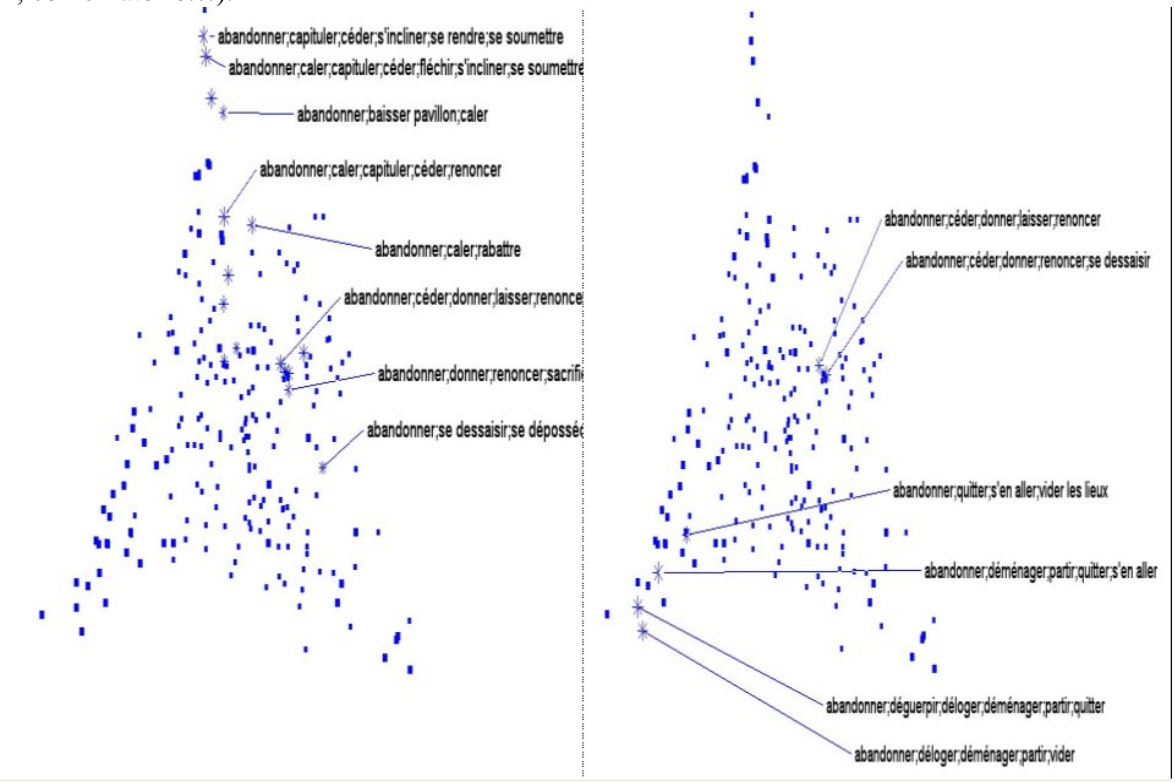


Figure 5. Projection de définitions du TLFI sur l'espace sémantique de *abandonner*

8 Conclusion

Nous avons présenté ici un modèle de représentation continu du sens. Ce modèle présente l'avantage d'être adapté à la fois à des utilisations humaines et automatiques. Le choix de la relation de synonymie comme accès aux informations sémantiques lexicales s'est révélé pertinent pour la caractérisation sémantique des polysèmes. La question qui se pose désormais est celle de l'enrichissement et de la structuration de la ressource sémantique ainsi obtenue. Nous ne rendons compte pour l'instant que de certains aspects de la structure lexicale, puisque nous ne l'avons abordée que par le biais des relations de synonymie. Cette vue, partielle, est à compléter par l'étude d'autres relations lexicales qu'elles soient sémantiques (antonymie, hyperonymie...) ou non (dérivation, suites syntaxiques : adjectif – nom, verbe – adverbe..., rapports syntaxiques : verbe – nom (sujet), verbe – nom (objet)...). Un pas vers dans cette direction a déjà été fait avec la construction d'espaces distributionnels pour étudier la polysémie verbale (Jacquet & Venant, 2005). La prochaine phase consistera à tenter d'explorer d'autres graphes lexicaux tout en travaillant à stabiliser les modes de construction et d'exploration des espaces globaux. Les pistes envisagées, clustering sur les cliques, ou interaction avec d'autres modèles sémantiques semblent prometteuses.

References

Croft, William & D.Allan, Cruse. 2004. *Cognitive Linguistics*, Cambridge University Press.

Bartning, Inge. & Michèle, Noailly. 1993 du relationnel au qualificatif : flux et reflux, *L'information Grammaticale*, 58, L'adjectif (M. Noailly ed.).

Borodina, M.A. 1963. L'adjectif et les rapports entre sémantique et grammaire en français moderne., dans *Le Français Moderne*, XXXI-3, p. 193-198., 1963

Daille, Béatrice. 2001. « L'identification en corpus d'adjectifs relationnels : une piste pour l'extraction automatique de terminologie », *TAL, Volume 42 Lexiques sémantiques*

Falk, Ingrid, Claire, Gardent, Jacquey, Evelyne & Fabienne, Venant (à paraître). A method for grouping synonyms, Lexicography in the 21st century: new applications, new challenges. Louvain la neuve.

François, Jacques & Morgane, Sénéchal. 2004. Le sémantisme propre des cadres prédicatifs et la polysémie des verbes de parole, actes du colloque *La prédication,* Aix-en-Provence.

Goes, Jan. 1999. *L'adjectif entre nom et verbe,* Paris – Louvain –La - Neuve, Duclot.

Jacquet, Guillaume. 2003. Polysémie verbale et construction syntaxique : étude sur le verbe jouer, *Actes de la 10ème conférence annuelle sur le Traitement Automatique des Langues (RECITAL 2003)*, Batz-sur-mer.

Jacquet, Guillaume & Fabienne, Venant. 2005. Construction automatique de classes de sélection distributionnelle, Actes de la 10ème conférence annuelle sur le Traitement Automatique des Langues, TALN 05. Dourdan.

Jacquet, Guillaume, Fabienne, Venant & Bernard Victorri. 2005. Polysémie lexicale, dans ENJALBERT P. *Sémantique et traitement automatique du langage naturel*, Hermes.

Jayez, Jacques, 2008. Quel rôle pour les facettes ? *Langage,* volume 172.

Kleiber, Georges. 1999. *Problèmes de sémantique - la polysémie en questions*, Presses universitaires du Septentrion, Villeneuve d'Ascq

Noailly, Michèle. 1999. *L'adjectif en français moderne*, Paris, Ophrys.

Ploux, Sabine & Bernard Victorri. 1998. Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes, *Traitement automatique des langues*, 39/1, p.161-182, 1998

Polguère Alain. 2002. Le sens linguistique peut-il être visualisé ? In D. Lagorgette et P. Larrivée (dir.) : *Représentations du sens linguistique*, coll. "Lincom Studies in Theoretical Linguistics", 25, Munich : Lincom Europa, 89-103

Pustejosky, James. 1995. *The generative lexicon*, Cambridge, MIT Press, 1995.

Romero, Clara. 2004. « Les adjectifs intensifs », *François J. l'adjectif en français et à travers les langues*, Presses Universitaires de Caen, 2004.

Venant, Fabienne. 2004. Polysémie et calcul dynamique du sens, *Le poids des mots, actes des JADT 04*. Louvain La neuve.

Venant, Fabienne. 2006. Représentation et calcul dynamique : exploration du lexique adjectival du français, mémoire de doctorat de l'EHESS.

Venant, Fabienne. 2008. Représentation géométrique et calcul dynamique du sens lexical : application à la polysémie de livre. Langages, 172.

Victorri, Bernard. 1997. La polysémie : un artefact de la linguistique ? *Revue de Sémantique et de Pragmatique*, 2, p. 41-62.

Victorri, Bernard & Catherine, Fuchs. 1996. *La polysémie, construction dynamique du sens*, Paris, Hermès.

Description lexicographique des lexies dénotant des animaux dans un *Dictionnaire explicatif et combinatoire*

David Wilton
Dalhousie University, Halifax, NS
dave__wilton@hotmail.com

Résumé

Cet article propose la structure de l'article de dictionnaire des lexies dénotant des animaux et des composantes standard de leurs définitions dans un dictionnaire du type *Dictionnaire explicatif et combinatoire*. Des problèmes de distinction entre les connaissances linguistiques et encyclopédiques sont également traités. Nous discutons de la différenciation entre la connotation lexicographique d'une lexie L et une composante de la définition de L. Quelques dérivations et collocations des lexies dénotant des animaux qui mettent en jeu les connotations de ces dernières sont présentées, ainsi que la structure des vocables correspondants.

1 Introduction

Dans cet article nous présentons la description de quelques lexies dénotant des animaux selon la lexicologie explicative et combinatoire (Mel'čuk *et al*., 1995 et Mel'čuk, 2006). Cette description se fait dans le cadre du projet de dictionnaire *Dire autrement* (Milićević & Hamel 2007), un dictionnaire électronique d'apprentissage du français langue seconde pour le niveau intermédiaire-avancé.

Il n'y a pas beaucoup de descriptions de lexies dénotant des animaux dans notre cadre théorique, mis à part *Lexique actif du français* (Mel'čuk et Polguère, 2007) et *Le manuel électronique d'apprentissage du lexique* de Apresjan *et al.* (2007). En dehors de notre cadre, on peut mentionner, notamment, Wierzbicka (1985 : 146-257).

Nous nous concentrons sur un corpus de lexies décrites dans le cadre de notre mémoire de maîtrise (Wilton, *en préparation*). Ce corpus est constitué de 17 vocables (= mots polysémiques) dont la lexie de base dénote un animal, soit de 62 lexies. Nous ferons également référence à quelques autres lexies qui ne proviennent pas de notre corpus.

Plus particulièrement, nous nous intéressons aux questions suivantes :
1) La structure de la définition d'une lexie dénotant un animal [= L_{animal}] ; identifier les blocs standard et les difficultés reliées à cela : distinctions entre le savoir linguistique vs. non linguistique, distinctions entre la connotation d'une lexie et une composante de la définition de cette dernière.
2) Le statut de la connotation dans un *Dictionnaire explicatif et combinatoire* : le contenu et la forme de la zone de la connotation dans l'article de dictionnaire d'une lexie.
3) La structure du vocable dont la lexie de base est L_{animal} : les liens de polysémie entre les lexies du vocable, surtout des liens mettant en jeu la connotation.
4) Les lexies dérivées à partir des lexies qui dénotent des animaux et mettant en jeu les connotations de ces dernières.
5) Les collocations formées à partir des connotations des lexies dénotant des animaux.

Commençons par présenter la structure d'un article de dictionnaire explicatif et combinatoire [= DEC]. Un article de dictionnaire de type DEC comporte les quatre zones principales suivantes :

- Zone sémantique, constituée de deux sous-zones : la définition de la lexie-vedette L, c'est-à-dire la décomposition du sens de L en termes de sens plus simples que celui de L [déf], et les connotations de L [ct].
- Zone syntaxique [tr] : le tableau de régime de L (*grosso modo*, le cadre de sous-catégorisation de L).
- Zone de combinatoire lexicale [fl]: la cooccurrence lexicale restreinte de L, c'est-à-dire des collocations et des dérivations de L, décrite en termes de fonctions lexicales (= outils formels pour la description des relations lexicales ; voir, par exemple, Wanner (éd.), (1996).
- Zone d'exemples [ex] : exemples de phrases mettant en jeu la lexie-vedette.[1]

[1] Nos exemples sont tirés d'une recherche sur *Google*.

Voici l'article de dictionnaire de la lexie MOUTON#I.1, la lexie de base du vocable correspondant (provenant de notre corpus) :

(1) [déf] mouton DE L'individu X =
animal domestique
de taille moyenne
à pelage frisé
que X élève pour
sa laine, sa viande et sa peau
[ct] 'douceur' [*doux comme un mouton*]
'docilité' [*docile comme un mouton*]
'crédulité/naïveté' [*crédule/naïf comme un mouton*] (voir MOUTON#II.1a)
[tr] X = I = de N, A-poss [*moutons de Jean* ; *ses moutons*]
[fl] {A_0 = adjectif relationnel} moutonnier, ovin
{Son = M. émet un cri} bêler
{S_0Son = cri du M.} bêlement
{onomatopée} bê
{mâle du M.} bélier
{femelle du M.} brebis
{jeune du M.} agneau
{Mult = ensemble de M.} troupeau
{$Real_1$ = ce que X est censé faire} garder [ART ~s]
{S_1Real_1 = celui qui garde les M.} berger
{S_{loc} = lieu où X garde les M.} bergerie
{pelage du M.} toison
{f_1= X enlever la toison du M.} tondre
{$S_0(f_1)$} tonte [des ~s]
{f_2 = mettre bas des M.} agneler || femelle du M.
{$S_0(f_2)$} agnelage || femelle du M.
[ex] *Voici un fermier qui tond son mouton à l'aide d'une tondeuse électrique.*

Nos définitions sont des **définitions didactisées** (≈ adaptées aux besoins spécifiques d'apprenants), comme celles du *Dire Autrement*, proposée dans Milićević, 2008. Pour ce qui est des fonctions lexicales, nous les indiquons en versions classique et « vulgarisée » (Popovic, 2003), cette dernière étant utilisée dans d'autres DEC à visée pédagogique. La version vulgarisée d'une fonction lexicale [= FL] est un type de paraphrase du sens de la FL. Ainsi, par exemple, la FL A_0 (≈ la dérivation syntaxique adjectivale du mot-clé) est présentée avec la vulgarisation : adjectif relationnel.

Selon le principe d'adéquation (Iordanskaja & Mel'čuk, 1984 : 27), chaque composante de la définition d'une lexie L doit être nécessaire et l'ensemble des composantes doit être suffisant pour couvrir tous les emplois de L. Cependant, comme nous sommes dans le contexte d'un dictionnaire d'apprentissage, nous nous permettons certaines simplifications dans les définitions et lorsque nous jugeons de l'adéquation d'une définition, nous prenons la reconnaissance des lexies correspondantes comme facteur central (Apresjan *et al.*, 2007). Une composante sémantique jugée non essentielle pour la reconnaissance de la lexie n'est incluse dans la définition que si elle assure le lien avec une autre lexie du vocable, avec une dérivation ou avec des collocations.

Dans ce qui suit, nous nous concentrons sur la zone sémantique des articles des lexies dénotant les animaux. Le reste de l'article se structure comme suit : la définition et les problèmes reliés font l'objet de la section 2 ; plus particulièrement, il y sera question de la distinction entre les connaissances linguistiques et encyclopédiques et de la distinction entre les composantes de la définition de L et les connotations de L. La connotation est traitée dans la section 3 ; nous y présenterons les évidences linguistiques nécessaires pour postuler une connotation d'une lexie L. La section 4 porte sur la structure des vocables correspondants ; nous y présenterons les schémas de polysémie entre les lexies d'un vocable dont la lexie de base est L_{animal}. La conclusion constitue la section 5.

2 Définitions des lexies dénotant des animaux

Dans cette section nous traiterons des trois points suivants : la structure de la définition d'une lexie L_{animal} (2.1) ; la distinction entre les connaissances linguistiques et les connaissances encyclopédiques (2.2) ; et la distinction entre les informations qui doivent apparaître en tant que composantes de la définition de L_{animal} et celles qui doivent apparaître dans la zone de la connotation de L_{animal} (2.3).

2.1 Structure de la définition

La définition d'une lexie L comporte deux parties : le **défini** et le **définissant**. Le défini est la **forme propositionnelle** de L, une expression contenant L et ses **actants sémantiques** (≈ participants obligatoires de la situation décrite par L). Le définissant est, comme nous l'avons dit plus haut, la décomposition du sens de L en termes de sens plus simples que celui de L[2] ; cf. :

(2) ÂNE#I
[défini] âne DE L'individu X =
[définissant] animal domestique
de grande taille
avec de longues oreilles
que X élève pour le transport

Une lexie L_{animal} peut correspondre à un **nom d'objet sémantique** ou à un **quasi-prédicat sémantique**. Un (nom d')objet sémantique est un sens complet et non liant dans le sens où il n'exige pas que l'on exprime auprès de lui d'autres sens ; autrement dit, une lexie qui correspond à un objet sémantique n'a pas d'actants sémantiques. Les lexies dénotant des animaux sauvages, comme CARIBOU#I.1 et TAUPE#I.1, sont des objets sémantiques. Une lexie qui correspond à un quasi-prédicat dénote une entité qui est liée à une situation particulière de fonctionnement ou d'utilisation, et c'est de cette situation qu'elle tire ses actants sémantiques. La plupart des lexies dénotant des animaux domestiques, telles que MOUTON#I.1 et PORC#I.1, correspondent à des quasi-prédicats et ont un actant : l'individu X qui s'occupe d'eux ou les utilise d'une certaine façon (sur les quasi-prédicats, voir Mel'čuk et Polguère, 2008).

Une définition peut contenir des **blocs standard** : (configurations de) composantes qui apparaissent dans les définitions de beaucoup de lexies appartenant à un même champ sémantique. C'est la pratique courante dans la lexicographie théorique de décrire le sens des lexies à partir de telles composantes. Pour les blocs standard dans les définitions des lexies dénotant des animaux, voir Wierzbicka, 1985 et Apresjan *et al.*, 2007.

Nous proposons les 6 blocs standard suivants pour les définitions des lexies dénotant des animaux. (Bien entendu, ce ne sont pas tous les blocs standard qui apparaissent dans chaque définition.)

1) Composante centrale

La composante centrale du sens d'une lexie L correspond à une **étiquette sémantique** (Milićević, 1997) qui identifie le sens générique de L. Une étiquette sémantique définit une classe de lexies – les lexies d'une classe partagent des propriétés de cooccurrence libre et restreinte ; on dit alors que ces lexies héritent des propriétés de leur étiquette sémantique. Nous nous sommes servi des étiquettes sémantiques organisées en une hiérarchie, reprises de Grizolle (2003) :

[2] Une lexie L_1 et considérée plus simple qu'une lexie L_2 lorsque L_1 peut apparaître dans la définition de L_2 mais non pas vice-versa. Prenons, par exemple, les lexies REGARDER et DÉVISAGER. La lexie REGARDER doit se trouver dans la définition de DÉVISAGER, mais DÉVISAGER ne peut pas se trouver dans la définition de REGARDER.

animal	
|-------> oiseau	AUTRUCHE, COLIBRI, PERROQUET
|-------> animal sauvage	CARIBOU, LOUP, TAUPE
|-------> animal domestique	ÂNE, CHEVAL, LAPIN
| |-------> de compagnie	CHAT, CHIEN, HAMSTER
| |-------> d'élevage	MOUTON, PORC, POULE

2) Taille

Les trois composantes suivantes nous ont servi pour décrire la taille dans nos définitions : 'de petite taille', 'de taille moyenne' et 'de grande taille'. Lorsqu'une lexie L dénote un animal dont la taille adulte est considérablement variée, nous avons omis la composante de la taille, par exemple, dans la définition de CHIEN#I. De même, nous avons omis la composante de la taille lorsqu'une lexie L_{animal} désigne l'espèce correspondante, par exemple OISEAU#I. La taille dans les cas ci-dessus est souvent trop variée pour justifier sa mention dans les définitions. Cependant, ce bloc est utile lorsque nous décrivons une lexie L_{animal} d'une taille restreinte ; par exemple, la lexie CHIHUAHUA est décrite en tant qu'un 'CHIEN#I de petite taille…'. Ainsi, la taille de l'animal est toujours décrite par rapport à l'homme et non pas par rapport à l'espèce correspondante. Cette façon de faire n'oblige pas le lecteur à connaître toutes les « races » de l'espèce correspondante, ainsi que toutes les limites de la taille de cette dernière, pour comprendre, par exemple, la composante 'de taille moyenne'. Voici quelques exemples des composantes de la taille :

'de petite taille'	'de taille moyenne'	'de grande taille'	'---'
SOURIS	TERRE-NEUVE#II	ÂNE	OISEAU
LAPIN	PORC	CHEVAL	REPTILE
CHIHUAHUA	MOUTON	ÉLÉPHANT	CHIEN

3) Aspect physique

Il existe deux blocs standard pour la description de l'aspect physique des animaux : 'parties du corps' et 'caractéristiques de la peau, du pelage ou des poils'.

Parties du corps :

KANGOUROU 'deux pattes'
ÂNE 'longues oreilles'
CHEVAL 'longue crinière'
LION 'crinière touffue'

La définition de KANGOUROU contient la composante 'deux pattes' puisque sa composante centrale est animal sauvage. La définition de animal sauvage inclut l'information 'animal à quatre pattes', par conséquent cet aspect physique du kangourou constitue une exception. Puisque l'âne et le cheval se ressemblent considérablement, nous avons ajouté 'longues oreilles' à la définition de la lexie ÂNE pour assurer la différenciation entre les deux.

Caractéristiques de la peau, du pelage, des poils :

TAUPE 'poils de couleur grise aux reflets bruns' [la couleur TAUPE#II.1]
ÉLÉPHANT 'peau rugueuse' [avoir une peau d'éléphant]
TIGRE 'pelage rayé' [TIGRÉ]
MOUTON 'pelage frisé' [TOISON]

Les composantes ci-dessus sont nécessaires à la fois pour la reconnaissance de la lexie décrite L et pour assurer le pont sémantique entre L et une autre lexie du même vocable (Mel'čuk *et al*., 1995 : 157) ou pour faire un lien collocationnel ou dérivationnel. Par exemple, la composante 'de couleur grise à reflets bruns' dans la définition de TAUPE#I est nécessaire pour faire le pont avec la lexie TAUPE#II.1, qui désigne la couleur correspondant à la peau de TAUPE#I ; la 'peau rugueuse' de l'éléphant est nécessaire pour décrire la collocation *avoir une peau d'éléphant* || 'avoir une personnalité [peau#II.1] peu sensible' ; le 'pelage rayé' du tigre décrit la dérivation adjectivale TIGRÉ || '[X] tigré = qui est marqué de bandes foncées' ; finalement, le 'pelage frisé' du mouton est la décomposition approximative de TOISON || 'pelage laineux et frisé'.

4) Le comportement

Le comportement des animaux est décrit à partir de trois composantes : ‘qui se nourrit de [nourriture]’ / ‘qui se nourrit par [partie du corps]’, ‘qui se déplace en [manière de déplacement]’ et ‘qui fait [activité typique]’.

Façon de se nourrir :

LAPIN ‘qui se nourrit de légumes’
OISEAU ‘qui se nourrit par un bec’
ANIMAL SAUVAGE/DOMESTIQUE ‘qui se nourrit par une gueule’
ÉLÉPHANT ‘qui se nourrit avec une trompe’

La composante ‘qui se nourrit de légumes’ dans la définition de LAPIN#I.1 est nécessaire pour décrire la collocation *manger comme un lapin* (≈ ne se nourrir que de salades).

Façon de se déplacer :

OISEAU ‘qui se déplace en volant’
SERPENT ‘qui se déplace en rampant’

Activités typiques :

TAUPE ‘qui creuse des tunnels souterrains’
ARAIGNÉE ‘qui construit des toiles’

La composante ‘creuser des tunnels souterrains’ est nécessaire pour assurer le pont sémantique avec la TAUPE#II.2 ‘dispositif utilisé par l’individu X pour creuser des tunnels souterrains’ (*J’ai utilisé la taupe#II.2 pour faire le trou*).

5) La relation avec l’homme

Les six composantes suivantes décrivent la relation entre l’homme et l’animal. Certaines de ces composantes sont héritées de la composante centrale, par exemple la composante ‘que X garde pour sa compagnie’ est héritée de l’étiquette sémantique animal de compagnie par les lexies correspondantes.

Que X garde pour sa compagnie :

CHIHUAHUA
PERRUCHE

Que X garde pour ses compétences :

CHAT ‘chasser les souris’
CHEVAL ‘travaux agricoles’, ‘transport’

Que X élève pour en tirer des produits :

MOUTON ‘viande’, ‘laine’, ‘peau’
POULE ‘viande’, ‘œufs’

Que X chasse pour en tirer des produits :

CARIBOU ‘viande’, ‘peau’
ÉLÉPHANT ‘ivoire’

Que X considère comme nuisible :

RAT ‘qui peut transmettre des maladies’
TAUPE ‘qui laisse des monticules de terre dans les jardins’

Que X considère comme dangereux :

TIGRE ‘qui peut tuer l’homme’
LOUP ‘qui peut tuer l’homme’

6) Habitat

Cette composante n’est pas incluse dans la définition des animaux qui sont répandus sur tout le globe et n’est indiquée que pour ceux vivant dans des régions spécifiques.

LION ‘vivant en Afrique et en Asie’
TIGRE ‘vivant en Asie’
CARIBOU ‘vivant dans des régions polaires de l’Amérique du Nord’
KIWI ‘vivant en Nouvelle-Zélande’

Cette composante assure dans certains cas le pont sémantique, comme dans le cas des lexies CARIBOU#II ‘Canadien [vivant au même endroit que le CARIBOU#I.1]’ et KIWI#II (en anglais) ‘Néo-zélandais [vivant au même endroit que le KIWI#I]’.

Voici les définitions de deux lexies de notre corpus avec indication des blocs standard :

(3) CARIBOU#I.1
[composante centrale] animal sauvage
[taille] de grande taille
[aspect physique] avec de grands bois#III.4
[habitat] qui vit dans des régions polaires de l'Amérique du Nord
[relation avec l'homme] et que l'homme chasse pour sa viande et sa peau

(4) TAUPE#I.1
[composante centrale] animal sauvage
[taille] de petite taille
[aspects physiques] avec un museau long et pointu
aux poils gris aux reflets bruns
[comportement] qui creuse des tunnels souterrains
dans lesquels il vit
[relation avec l'homme] et que l'on considère comme nuisible
parce qu'il laisse des monticules de terre dans les jardins.

2.2 Distinction entre les connaissances linguistiques vs. encyclopédiques

Il est parfois difficile de préciser la frontière entre les connaissances linguistiques (qui font partie de la description lexicographique) et les connaissances encyclopédiques (qui ne doivent pas faire partie d'une telle description). Wierzbicka (1985 : 200) insiste sur le fait qu'un dictionnaire doit décrire les lexies en se basant sur les connaissances « naïves » sur le monde, car ce sont les connaissances de ce type qui se reflètent dans le lexique. La terminologie du spécialiste, qui apparaît dans beaucoup de définitions des dictionnaires existants, ne fait pas partie des connaissances d'un locuteur « normal » et représente les connaissances encyclopédiques, plutôt que linguistiques.

La définition de CHEVAL de l'*Oxford English Dictionnary*, reprise de Wierzbicka (1992 : 49), illustre l'incapacité de certains dictionnaires de différencier entre ces deux types de connaissances :

(5) HORSE : a solid-hoofed perissodactyl quadruped.

Aucune composante de cette description n'aide à la reconnaissance de la lexie. Le vocabulaire utilisé ne s'emploie pas dans le langage courant et il est fort probable que le lecteur qui comprend la composante 'perissodactyl' comprend également la lexie HORSE.

Les types de composantes problématiques comme ceux dans la définition ci-dessus se trouvent également dans les définitions des dictionnaires du français ; cf. la définition d'ARAIGNÉE tirée du *Petit Robert* :

(6) ARAIGNÉE : **arachnide** (**aranéides**) dont la taille peut varier d'une fraction de millimètre à vingt-cinq centimètres environ, muni de crochets à venin et de **glandes séricigènes**.

Les composantes en gras représentent les composantes que nous considérons comme problématiques. Les composantes 'arachnide', 'aranéides' et 'glandes séricigènes' ne constituent pas, à notre avis, une décomposition du sens d'ARAIGNÉE selon un locuteur ordinaire. De plus, la composante décrivant la taille représente plus que la moitié de la définition et tout cela pour conclure avec la composante 'environ' ; une encyclopédie fournit cette information et de manière encore plus précise. Voici la définition que nous proposons pour la lexie ARAIGNÉE :

(7) ARAIGNÉE : petit animal
à huit pattes
qui construit des toiles pour attraper les insectes dont il se nourrit
et dont certaines espèces ont du venin.

2.3 Connotation vs. composante de la définition

La connotation lexicographique d'une lexie L est une composante sémantique associée par la langue au référent de L, mais qui ne fait pas partie de la définition de L. Par exemple, la composante sémantique 'jalousie' est une connotation de la lexie TIGRE#I. Cette composante ne fait pas partie de la définition de la lexie TIGRE#I, mais se trouve plutôt dans une zone séparée, intitulée la **zone de connotation** [= `ct`]. Une connotation lexicographique d'une lexie L doit figurer dans l'article de dictionnaire de L si elle est nécessaire pour faire un lien vers une collocation de L, une dérivation de L ou une autre lexie du même vocable que L.

Il n'est pas toujours facile de distinguer entre la connotation et une composante de la définition. Par exemple, est-ce que la notion 'stupidité' devrait faire partie de la définition de la lexie ÂNE#I 'animal domestique…' ? Il existe des tests pour déterminer si un sens est une composante de la définition d'une lexie L ou une connotation de L (Iordanskaja & Mel'čuk, 1984 : 33-40). Nous en présentons un : le test d'antonymie.

Le test d'antonymie est utile lorsqu'un modificateur d'une lexie L a un antonyme. Prenons une lexie L avec une connotation hypothétique (C). Si l'ajout d'un antonyme de (C) à L produit une contradiction, (C) n'est pas une connotation mais une composante de la définition de L, autrement (C) est une connotation. Voici un exemple :

(8) a. ÂNE#I 'animal…' `ct` = 'stupidité'
b. ÂNE#II 'individu stupide…'

La phrase *L'âne#I de mon père est intelligent* est sémantiquement cohérente, ce qui indique que la composante 'stupide' ne doit pas figurer dans la définition d'ÂNE#I. Cependant, la phrase $^{\#}$*Mon frère David est un âne#II intelligent* est contradictoire, alors la composante 'stupidité' doit paraître dans la définition d'ÂNE#II.

En principe, une caractéristique objective d'une lexie va plutôt dans la définition, alors qu'une caractéristique subjective d'une lexie représente plutôt une connotation. La définition de LAPIN#I.1 comporte la composante 'dents longues', qui fait le lien avec la collocation DENTS *de lapin*. En anglais, la composante 'long ears' dans la définition de RABBIT#I.1 fait le lien avec la lexie ⌜RABBIT EARS⌝ (= 'antenne portative de télévision ayant la forme des oreilles d'un lapin').

3 Connotation

La présente section vise à illustrer les évidences linguistiques nécessaires pour postuler une connotation. La sous-section 3.1 porte sur les dérivations de L mettant en jeu les connotations de cette dernière et la sous-section 3.2 traite des collocations mettant en jeu les connotations.

3.1 Dérivations mettant en jeu les connotations

Les dérivations mettant en jeu les connotations qui figurent dans notre corpus appartiennent à deux parties de discours : les dérivations nominales et les dérivations adjectivales. Il s'agit des lexies dérivées soit à partir d'une lexie qui dénote un animal, soit à partir d'une lexie qui dénote un individu ayant une certaine caractéristique (liée par connotation à la lexie dénotant l'animal). Les lexies dérivées dénotent des actes, des énoncés, et des caractéristiques des individus ou des faits.

(9) ÂNE#II 'individu X stupide [comme si X était un ÂNE#I]' `ct` = 'stupidité'
Arrête de faire l'âne.
ÂNERIE#1a 'caractère stupide d'un individu X [comme si X était un ÂNE#II]'
L'ânerie de Jean est proverbiale.
ÂNERIE#1b 'caractère stupide d'un fait X [comme si X était attribuable à un ÂNE#II]'
Ceci montre l'ânerie de la formation classique qui demande d'optimiser un code pour faire joli et montrer qu'on a compris.

ÂNERIE#2a ‘acte stupide de Y-er
fait par l’individu X [comme si X était un ÂNE#II]’
Il faut éviter de commettre l’ânerie de s’inscrire à la Sacem.
ÂNERIE#2b ‘énoncé stupide
de l’individu X / au sujet de Y [comme si X était un ÂNE#II]’
À mon avis, tu n’as pas réfléchi beaucoup pour pondre une pareille ânerie.

(10) LAPIN#I.1 ‘animal …’ `ct` = ‘prolificité’
Le lapin est élevé pour sa chair.
LAPINISME ‘fécondité excessive de X’ [comme si X était un LAPIN#I.1]’
C’est le lapinisme de certaines populations qui fait qu’il y a de plus en plus de crève-la-faim dans le monde.

(11) MOUTON#II.1 ‘[individu X] crédule [comme si X était un MOUTON#I.1]’ `ct` = ‘crédulité’
Et oui si tu suis bêtement la mode tu es mouton.
MOUTONNERIE ‘caractère crédule d’un individu X
qui se manifeste en acte Y [comme si X était MOUTON#II.1]’
Le peuple français est coupable de sa propre stupidité, de sa moutonnerie.
MOUTONNIER#Iia ‘[individu X] qui se comporte comme un MOUTON#II’
C’est un public moutonnier qui se laisse docilement conduire vers l’ennui.
Ce film est un vibrant hommage à ceux qui se sont rebellés pour ne pas tomber dans le piège de la moutonnerie de suivre un chef sans foi ni loi.
MOUTONNIER#IIb ‘[fait X] dans lequel se manifeste la MOUTONNERIE’
C’est une conséquence de l'attitude moutonnière des investisseurs.

(12) PORC#II ‘individu X sale [comme si X était un PORC#I.1]’ `ct` = ‘saleté’
Je suis un vrai porc car je pète sans retenu.
PORCHERIE#II ‘endroit très sale [comme si c’était l’endroit où habite le PORC#I.1]
Je voudrais aussi ranger la porcherie de mon fils.

3.2 Collocations mettant en jeu les connotations

Il existe dans le cas de presque toutes les lexies de notre corpus qui dénotent des animaux, des collocations modélisées par les fonctions lexicales `Magn` ‘intense’/‘très’, `AntiMagn` ‘peu.intense’/‘peu’, `Bon` ‘positif (selon le locuteur)’ ou `AntiBon` ‘négatif (selon le locuteur)’. Nous avons également repéré une collocation modélisée par la fonction lexicale `AntiVer` ‘de façon qui n’est pas telle qu’elle devrait être’. La collocation a deux formes : « *comme* ART L_{animal} » ou « *de* L_{animal} ».

	Lexie	Connotation	Collocation
(13)	ÂNE	‘travailleur’	comme un âne [= `Magn`(TRAVAILLER)]
(14)	ÉLÉPHANT	‘rancune’	comme un éléphant [= `Magn`(RANCUNIER)]
(15)	LAPIN	‘rapidité’	comme un lapin [= `Magn`(DÉTALER)]
(16)	LION	‘courage’	comme un lion [= `Magn`(COURAGEUX)]
(17)	MOUTON	‘crédulité’	comme un mouton [`AntiVer`(SUIVRE)]
(18)	PORC	‘saleté’	comme un porc [= `AntiBon`(MANGER)]
(19)	RAT	‘avarice’	comme un rat [= `Magn`(AVARE)]
(20)	TIGRE	‘souplesse de mouvement’	comme un tigre [= `Magn`(LESTE)]

Dans notre corpus de 62 lexies, nous avons trouvé 38 `Magn`, 3 `AntiMagn`, 1 `Bon`, 3 `AntiBon` et 2 `AntiVer`. Notons que nous avons trouvé plusieurs collocations où l’existence d’une connotation (qui en serait la « source ») n’est pas évidente. Par exemple, la collocation *travailler comme un âne* est décrite en tant que `Magn`(TRAVAILLER) ayant la connotation ‘travailleur’ comme source. Peut-être pourrait-on également décrire cette collocation en tant que `AntiVer`(TRAVAILLER) ayant la connotation ‘stupidité’ comme source, puisque cette collocation peut aussi exprimer la notion de travailler sans être récompensé.

4 Structure des vocables

Dans les vocables dont la lexie de base est L_{animal}, certaines lexies du vocable sont reliées par métonymie pour désigner la viande, la fourrure, la peau ou les poils qui proviennent de l'animal dénoté par la lexie de base. D'autres lexies du vocable sont reliées à la lexie de base L par métaphore pour désigner des individus et des faits, le pont sémantique étant soit une composante de la définition de L (indiquée en gras dans les exemples ci-dessous), soit une connotation de cette dernière. Nous avons inclus seulement assez de lexies pour illustrer les types de liens existant entre les lexies dont la lexie de base est L_{animal}.

(21) PORC#I.1 'animal' `ct` = 'saleté'
PORC#I.2 'viande du PORC#I.1'
PORC#I.3 'peau du PORC#I.1'
PORC#II 'individu sale [comme s'il était un PORC#I.1]

(22) PORCHERIE#I.1 'endroit où X garde les PORCS#I.1' `ct` = saleté
PORCHERIE#I.2 'entreprise qui s'occupe de l'élevage des PORCS#I.1
PORCHERIE#II 'endroit très sale [comme si c'était une PORCHERIE#I]

(23) MOUTON#I.1 'animal … à **pelage frisé** …' `ct` = 'crédulité'
MOUTON#I.2 'viande du MOUTON#I.1'
MOUTON#I.3 'peau du MOUTON#I.1'
MOUTON#II.1 '[individu X] qui est crédule [comme s'il était un MOUTON#I.1]'
MOUTON#II.3 'petit nuage blanc et floconneux
qui ressemble au **pelage frisé** du MOUTON#I.1.

(24) TAUPE#I.1 'animal de **couleur grise**… qui **creuse des tunnels souterrains**…'
TAUPE#I.2 'poils de la TAUPE#I.1'
TAUPE#II.1 'de **couleur grise**…'
TAUPE#II.2 'dispositif destiné à **creuser des tunnels**'
TAUPE#II.3 'espion… **infiltré dans le milieu**… [comme si elle était une TAUPE#I.1 dans des tunnels souterrains]

(25) TIGRE#I.1 'animal' `ct` = 'férocité'
TIGRE#I.2 'fourrure du TIGRE#I.1'
TIGRE#II 'homme féroce [comme s'il était un TIGRE#I.1]'

(26) TIGRESSE#I 'animal' `ct` = 'jalousie'
TIGRESSE#II 'femme jalouse [comme si elle était une TIGRESSE#I]'

(27) RAT#I 'animal' `ct` = 'avarice'
RAT#II.1 'individu avare [comme s'il était un RAT#I]'

5 Conclusion

Cet article s'est intéressé à la définition et à la connotation lexicographique des lexies dénotant des animaux selon la Lexicologie explicative et combinatoire. Nous avons proposé des blocs standard pour la définition des lexies dénotant des animaux, indiqué des évidences linguistiques nécessaires pour postuler une connotation et décrit les « patrons » de polysémie qui mettent en jeu les composantes des définitions et les connotations des lexies dénotant des animaux.

Notre article n'a pas abordé plusieurs sujets importants en ce qui concerne les lexies dénotant des animaux. Une des questions que nous avons laissées en suspens concerne le statut des termes d'affection de type *mon petit loup*. Une description détaillée d'une lexie L_{animal} nécessite que l'on définisse le statut de ces expressions. Il reste également à traiter les locutions, y compris les proverbes, qui sont liées aux lexies dénotant les animaux.

Remerciements

Mille mercis à Jasmina Milićević pour toute sa patience et tous ses conseils durant la réalisation de cet article.

Références

Apresjan *et al*. (2007). O komp'jutornom učebnike leksiki russkogo jazyka [Le manuel électronique d'apprentissage du lexique russe]. *Russkij jazyk v naučnom osveščenii*.

Grizolle, B. (2003). *Classification des fonctions lexicales non standard du DiCo. Lexies étiquetées animal et artefact.* Rapport de stage. Montréal : Observatoire de Linguistique Sens-Texte.

Iordanskaja, L. & Mel'čuk, I. (1984). Connotation en sémantique et lexicographie. In: Mel'čuk, I., Arbatchewsky-Jumarie, N., Elnitsky, L., Iordanskaja, L., Lessard, A. (1984): 33-40.

Mel'čuk, I. (2006). Explanatory-Combinatorial Dictionary. In: Sica, G. (ed.). *Open Problems in Linguistics and Lexicography*. Monza: Polimetrica Publisher; 225-355.

Mel'čuk, I. *et al.* (1988-1984-1992-1999. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques I-IV*. Montréal : Les presses de l'Université de Montréal.

Mel'čuk, I., Clas, A. & Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve : Duculot.

Mel'čuk, I. & Polguère, A. (2007). *Lexique actif du français. L'apprentissage du lexique fondé sur 20 000 dérivations sémantiques et collocations du français.* Bruxelles : de Boeck.

Mel'čuk, I. & Polguère, A. (2008). Prédicats et quasi-prédicats sémantiques dans une perspective lexicographique. *Revue de linguistique et didactique de langues,* Université de Grenoble, 2008, nº 35. [Site BLS du COURS]

Milićević, J. (1997). *Etiquettes sémantiques dans un dictionnaire formalisé du type Dictionnaire Explicatif et Combinatoire*. Mémoire. Montréal : l'Université de Montréal.

Milićević, J. (2008). Structure de la définition lexicographique dans un dictionnaire d'apprentissage explicatif et combinatoire. In : Bernal, E. & DeCesaris, J., eds., *Proceedings of the XIII EURALEX International Congres*. Barcelona, 15-19 July, 2008. Barcelona: University Institute for Applied Linguistics, Pompeu Fabra University: 551-561.

Milićević, J. & Hamel, M.-J. (2007). Un dictionnaire de reformulation pour les apprenants du français langue seconde. Dans: Chevalier, G. *et al*. (eds.), *Les apports de la sociolinguistique et de la linguistique à l'enseignement des langues en contexte plurilingue et pluridialectal*. PAMAPLA 29/Actes du 29e Colloque annuel de l'ALPA tenu à l'Université de Moncton, 4-5 nov. 2005, la Revue de l'Université de Moncton, Numéro hors série 2007; 145 à 167.

Popovic, S. (2003). *Paraphrasage des liens de fonctions lexicales*. Mémoire de maîtrise. Département de linguistique et de traduction, Université de Montréal.

Wanner, L., ed. (1996). *Lexical Functions in Lexicography and Natural Language Processing*. Amsterdam/Philadelphia: Benjamins.

Wierzbicka, A. (1985). *Lexicography and conceptual analysis*. Ann Arbor : Karoma.

Wierzbicka, A. (1992). What are the uses of theoretical lexicography? *Dictionaries : The Journal of the Dictionary Society of North America* 14: 44-78.

Wilton, D. (en préparation). *Description lexicographique des lexies dénotant des animaux dans un Dictionnaire explicatif et combinatoire*. Mémoire de maîtrise. Halifax : Dalhousie University.

A Target-oriented Case Study on Some Chinese Syntax-based Semantic Restrictions

Xiaohong Wu
Centre Tesniere, Faculty of Letters,
University of Franche-Comte, France
Faculty of Foreign Languages,
Minzu University of Qinghai Province,
China
wuxiaohongfr@yahoo.com.cn

Sylviane Cardey
Centre Tesniere, Faculty of Letters,
University of Franche-Comte, France
sylviane.cardey@univ-fcomte.fr

Abstract

Concerning the relationships between a verb and its complements much work has been done in recent years, especially those between a verb and its nominal and prepositional complements. However, another fundamental yet not deeply studied problem for the Chinese language in the application of machine translation (MT) is that the relationship between a verb and its complements of other grammatical categories, e.g. adjectives, still remains a big challenge. Such relationships are often related to certain particular syntactic structures and are thus bound by syntax-based semantic restrictions. In the case of Chinese linguistics, the ba-construction and its counterpart structure manifest exactly such a kind of phenomenon. In this paper we demonstrate the problems we encounter by illustrating the syntactic and semantic variations produced by these structures. We explain in detail how, if there is a lack of such syntactic and semantic information, the MT quality is influenced by generating badly constructed Chinese sentences which are neither acceptable nor grammatical. Since we apply controlled language (CL) technique as an aid to our work, we narrow our research to limited syntax and restricted semantics.

1 Introduction

In this paper we describe the work we do for the LiSe project[1], focusing on machine translation in the domains where safety is extremely crucial and the translation of such texts thus requires a high degree of accuracy. Due to the characteristics of the project, we start the research work first from the design and development of a controlled language (CL) in the hope that taking the advantages of CL we could greatly diminish the syntactic and semantic complexity and ambiguity from all levels for the languages with which we work with. We select three sub-domains for the CL applications: medical protocols, public warnings and certain emergency messages (for more information, see Wu, et al. 2006; Wu, 2005). Texts of this kind in our corpus are similar to those of users' manuals or practical instructions and thus are relatively easier to reduce in size or make structurally simpler and semantically less ambiguous. All the texts are translated from French into four target languages: Arabic, Chinese, English and Thai by human translators and verified by domain specialists. By doing so, we successfully build several parallel corpora for our language comparison and linguistic analysis, taking French as the source language (SL) and the others as the target languages (TL).

These parallel corpora are then distributed to different linguists of each language for re-editing, that is, to apply CL technique to make the sentences in the text shorter and simpler using the same rewriting guidelines. The re-edition of the texts is often done with the intervention of the domain specialists so as to make the sentences not only structurally simpler but also easier to understand. Except for English and

[1] LiSe (Linguistique et Sécurité), project funded by the ANR (French National Research Agency) contract: ANR-06-SECU-007.

French which are relatively more close to each other, the other three languages are linguistically quite far from these two and from each other. Without the help of CL techniques it would be undoubtedly a big challenge. In this paper, we present as an example only part of the work performed on three of the languages so as to demonstrate the problems produced by the linguistic discrepancies among them: Chinese, English and French. In addition, our presentation also centers on two basic Chinese structures – the ba-construction and the cohesive complement to show how problems arise if there is a change of the structure or the word order.

2 Language particularities and problems faced

Discussions on ba-construction have remained one of the most important topics in Chinese linguistics circles. The enormous number of results provided by an Internet search on the topic might give us an impression that there is no room for anyone to continue on this topic. It is true that in relation to the possible constructions of the structure, no matter from a syntactic or a semantic point of view, almost all facets of this construction have been explored both diachronically and synchronically. However, much of such linguistic research is found to be difficult to be integrated directly in the application of MT. Consequently, this frequently used Chinese syntactic construction is seen to be very badly translated into or, especially from other language(s). An obvious evidence for this bad performance is that the ba-construction is by itself a Chinese language specific phenomenon which finds no place in other languages. Besides, this superficially simple structure (ba + NP + VP) in fact covers almost all the types of Chinese phrasal constructions and also owns some semantically very complicated particularities according to the varied ways it is employed, of which the latter is still far beyond the description of present computational linguistic knowledge. Furthermore, the semantic content of this construction may vary if any change of the word order in the syntax happens. Therefore, the difficulty lies directly in the generation of the Chinese equivalents from other language(s) if the ba-construction is mandatory in interpreting the meanings conveyed in the source language (SL), such as in our case.

In our work we firstly build our own specific lexical information and grammar rules for the creation of this construction while transferring a French sentence into a Chinese one which needs this construction. Our analysis is based on the systemic or micro-systematic linguistic analysis theory which stresses the fact that a model ought to serve not just for the description of a precise phenomenon but also as a resource allowing solutions to other linguistic phenomena covered by the model as well as the linguistic phenomena not yet identified for which the model and/or others could use. The theory advocates the decomposition or recomposition of linguistic phenomena in order to analyze them better. For example, instead of listing all the elements which make up part of a language, it is better to try to classify, order, arrange or group them so as to define if certain of them can function as a complete independent system or as interrelated systems according to what ought to be demonstrated or solved. The theory is by itself based on set theory. When confronted with a problem or a precise phenomenon we try to choose the necessary elements and structure them in the form of a system. In this way, the problem can be manipulated as it is apprehensible. Generally, lexis, morphology, syntax and semantics cannot be separated when one deals with language phenomena. It is thus necessary to: identify the problem to be solved; construct the system or the systems which can be interrelated; describe then the problem using the system/s in order to solve the problem where necessary. Applications of the theory can involve different languages, different problems and different domains (Cardey, 2009).

While defining the syntactic rules both for the phrasal level and for the sentence level, we came across some difficulties caused by the language particularities even in some basic and simple structures. For example, imperative sentences take an important role in our corpus and this kind of sentences has to be translated into Chinese in most cases by generating a sentence containing the ba-construction. Therefore, we need to mark this information directly on the verbs which can be used in this structure, such as [± ba]. However, some verbs can be used in both ways, that is, either with the ba-construction or without it. It seems easy to add a rule indicating both options. The problem here is that this seemingly easy technique

is often jeopardized by some syntactic-semantic relationships inhabited within this structure, which is quite different from that of the English and French counterparts.

2.1 Language particularity – the ba-construction

Superficially, the ba-construction has a simple structure which can be represented as ba + NP + VP. Zhan (2004) classifies the types of both NP and VP compositions in the ba-construction from two aspects, one from the phrasal syntactic types of VP and the other from the semantic role types that the NP can take. He therefore classifies the syntactic construction of VP into 12 types and the semantic roles of the NP into 9 types. In his paper Zhan also offers the statistical figures calculated in relation to the ba-construction. He took the Modern Chinese Grammar Information Dictionary (published in 1997) as a resource for this work. According to his investigation, this dictionary lists 10,283 verbs among which 5,933 can take nominal complements. 1,803 verbs out of all these verbs can be used in ba-construction, which according to his further examination should be more than that as he found out that some verbs that otherwise can be used in this construction are not so marked. However, all these figures do not provide us with any immediate inspiration in that we have a limited list of verbs and syntax to worry about. Unlike European languages, the Chinese language does not possess a large vocabulary but it has a strong capacity for word formation and many words are occasionally formed or can be framed for a temporary use with added unpredictable semantic information. This might produce some degrees of influence for both the syntax and the semantics of a given sentence. This information is hard to catch and much of it is inevitably far beyond our present scope of research.

Generally speaking, only transitive verbs can be used in this construction. Therefore many sentences with a transitive verb can be transferred into a sentence with the ba-construction. That is, sentences that are of SVO sequence can be changed into ones with the ba-construction. When a sentence which has the surface structure of SVO is transferred into the ba-construction, some changes take place to the general structure. The object (usually a NP) of the verb is shifted to the front of the verb which makes the whole SVO structure possesses the surface structure of an SOV sequence, having the NP become the complement of the preposition ba.

Another fact is that ba as a preposition lacks semantic content by itself. However, it constrains the selection of verbs as we have mentioned. In other words, not all transitive verbs can be used in this construction. According to Ye, et al. (2007), the whole construction has "abstract meanings" as "disposal" or "causation" which are independent of the content words inhabiting it. "Therefore only transitive verbs that encode such meanings are permitted to appear in this construction". They further stress that 'a verb which is perfectly acceptable in the SVO structure but does not have the appropriate "disposal" or "causation" meaning would constitute a construction-based semantic violation if it is forced to appear in the ba-construction'. They refer to this as a construction-based semantic constraint which cannot be attributed to the word ba.

Another equally important phenomenon closely related to this structure is that after the object of the transitive verb has been shifted to the front and a ba-construction is formed, many verbs by themselves become semantically incomplete and thus require another complement to fill the slot left behind by the shifted object. This newly obtained complement is often made up of words of other grammatical categories or a phrase. This new complement often adds some more semantic or even pragmatic nuances to the verb and/or the sentence to indicate different levels of information associated with the action. Such nuances differ from those expressed in their SVO counterpart.

Interestingly, many such syntax-based semantic violations can be resolved by adding a post-posed complement(s) to the verb and result in the sentence being acceptable and grammatical. That is to say, if a verb cannot satisfy such a semantic requirement, it cannot be used in this construction, or we have to add more constituent(s) to this construction to satisfy this syntax-based semantic restriction.

Taking the examples of Ye et al. as ours, we can see that although the verb "欣赏 (close to 'enjoy/appreciate or view')" cannot be used in the ba-construction as in the example below:

a) *市民把名画欣赏 (LWS[2]: citizen ba famous picture enjoy/view).

we can make it canonical by adding a post-posed complement such as a verbal classifier. To make the sentence read more naturally we change the subject to a pronoun (我, I).

a.1 我把名画欣赏了一遍 (LWS: I ba famous picture enjoy/view once[3]; I enjoyed/viewed the famous pictures thoroughly).

By doing so, we have well satisfied the semantic 'need' for "causation".

In Chinese linguistics, this kind of postposed complement is often divided into two different kinds according to its syntactic behavior, of which one is called the cohesive complement and the other is called the compositive complement. Here lies our major problem: the interpretation and generation of the cohesive complement of the verb.

2.2 Language particularity – adverbial vs. cohesive complement – the problems faced

In all languages, for a certain meaning there often exist varied ways to express this, that is, one meaning can be expressed differently using different words and/or different syntactic structures (Milićević, 2006). Generally speaking, if a meaning is articulated using different ways, some semantic nuances might thus be produced. It is also possible that the general meaning is the same in the differently expressed sentences but that there might be some kind of syntactic restrictions for such different usages.

For example, in all these three languages: French, English and Chinese, some of the verb adjuncts (the constituents that modifies the verb) can appear both in front of the verb and after it. In Chinese, however, if a verb takes an adjunct that appears in front of the verb, we generally refer to it as the adverbial of the verb, which is in fact the most common position of the adverbial in Chinese. If the adjunct appears after the verb, it is considered as another kind of syntactic category hence forward called the cohesive/ compositive complement. This kind of verbal complement is different from the adverbial and such differences in the syntax leads to the dissimilarity of the grammatical function. Like the ba-construction this kind of constituent belongs to the category of language specific phenomena. Therefore, while translating a sentence into Chinese, we often have to "replenish" this kind of constituent to the translated sentence so as to make up for the semantic incompleteness required by the syntax.

Sometimes the same Chinese word (usually an adjective or an adverb) can appear in both positions (not at the same time). This results in two different syntactic constituents and often produces some change in the semantic emphasis. The option of the position in the phrasal structure depends not only on the sub-categorization scheme of the verb but also on the kind of syntactic structure where it is allowed to appear.

We look at the following two groups of VP, in which the same word appears in two different positions and demonstrates exactly such a situation despite the fact that we have just the same word but in a reversed order.

The change of the word order in the following examples does not produce much semantic difference but will be syntactically restricted if further used in the context of a sentence.

Group 1	躺 $_{v}$ 平 $_{adj}$	坐 $_{v}$ 端 $_{adj}$	放 $_{v}$ 平 $_{adj}$
Group 2	平 躺	端 坐	平 放
Translation	s'allonger sur le dos ?? lie on one's back	s'assoir ?? sit still	poser à plat ?? lay flat

Table 1 Structure variations caused by word order

As we have just mentioned, the reversed word order results in two different syntactic structures: while the first one is a verb plus an adjective forming a cohesive complement; the second can be regarded as an

[2] LWS: short for: literal word sequence

[3] "Once" here could also be interpreted as thoroughly or one after another

adverbial. Both of them can function as the predicate in a sentence. However, while the first group can be used as an independent sentence such as an imperative sentence, the second one when used in the similar way has to be supplemented with another constituent – a cohesive complement to indicate accordingly certain semantic needs as seen in the following example.

b) 请坐端 (Please sit still)。

However with the second choice, it has to be:

c) 请端坐在椅子上(LWS: please still sit on chair; Please sit still on the chair)。

This is instead of (d) which is ungrammatical:

d) *请端坐 (LWS: please still sit)。

This means that a verb having the cohesive complement tends to be semantically self-sufficient while a verb with the preposed adverbial in the same context is not. In other words, if we use the second group of words as the predicate in a sentence with the ba-construction, without the interference of a cohesive complement, it violates the syntax-based semantic restriction. This reflects the fact that the semantic restriction is not only required by the lexical unit but also by the syntactic structure. This also explains the reason why in some cases, this kind of violation might occur and/or such violations can be resolved by adding a cohesive complement. In Example (c), the prepositional phrase "on the chair" assures the semantic content of "location". As a result, there is no violation for (c) but there is so for (d).

It is clear that for a VP with only an adverbial, this tends to be more easily restricted by the syntactic-semantic relationships. However, simply reversing the position of the adverbial with the cohesive complement cannot always be the right solution as the actual situation is more complicated than that. This complication is often increased by the selection of the word(s) that can take the position of the adverbial or of the cohesive complement. To be brief, a word that can take the position of an adverbial does not necessarily mean that it can also take the position of a cohesive complement, and/or vice versa, though many of them might seemingly present almost the same syntactic or semantic resemblance. For certain of such phrases if we reverse the order, we will either create a structure which does not make any sense or a structure with quite different semantic content. See the examples in the following table.

Group 3	用(to use)坏(bad) : to wear out	倒(to pour)完 (over): to pour out	变(to turn)冷(cold): to turn cold
	*坏用	*完倒	*冷变
Group 4	擦 $_{v}$ 干 $_{adj}$ (to dry)	学 $_{v}$ 好 $_{adj}$ (to learn well)	弄 $_{v}$ 白 $_{adj}$ (to rub up or to whiten sth)
	干擦 (to rub/erase/clean with sth dry; or to rub in vain)	好学 (easy to learn)	白弄 (to do sth in vain)

Table 2 Examples of reversed word sequence

Words in Group 3 are exactly like those in Group 1: a verb plus an adjective. However, they cannot be used in a reversed way as shown in the row that follows. These reversed counterparts are not even the recognized Chinese words. The words in Group 4 when used in a reversed way do not have the same meaning at all, as the interpretations indicate.

Here the problem is manifest: if a system cannot distinguish an adverbial from a cohesive complement, or it lacks the capacity to deal with such phenomena, it will fail to generate the right sentences as is often seen in MT. This fact has caused us to start the study for the cohesive/compositive complement which until now appears to have been ignored.

Let us see a real example from our own MT system.

	French	English
e)	Laver bien toutes les bouteilles.	Wash well all the bottles.

Chinese

e.1 *把所有瓶子好洗。(LWS: ba all bottle well wash)

Using our own MT system, we first get a bad Chinese translation (see e.1) though it correctly matches the suitable lexical and syntactic equivalents that are predefined in the lexicon and in the grammar database. The MT system well generates the ba-construction as we have very efficient linguistic rules to deal with this structure. However, it makes a wrong judgment by taking "好" as the adverbial of the verb. In fact in this context "好" cannot take the position of the adverbial. If this word happens to appear in this position, it should be used in its duplication form, i.e. AA + V form, such as "好好 + V". For example, "把字好好写 (LWS: ba word well well write; Write carefully the words.)", "把话好好说 (LWS: ba speech/talk well well speak; Speak slowly/clearly/patiently <what you want to say>.)".

Yet, if "好" functions as the cohesive complement, it is well received. That is to say, when used in a "V + 好" structure, e.g. "把字写好 (LWS: ba word write well; Mind your handwritings <try to write beautifully>.)"; and "把话说好 (LWS: ba speech/talk speak well; Speak properly <what you want to say>.)". The problem here is that the change of the position produces the change of the meaning too, as we can see in the interpretations.

In addition, both of the just mentioned syntactic options are neither grammatically correct nor semantically exact in our case example as follows:

e.2 *把所有瓶子好好洗。(LWS: ba all bottle well well wash) → syntax-based semantic violation (it can be released by further adding a verbal classifier "一下" after the verb "wash".)

e.3 把所有瓶子洗好。 (LWS: ba all bottle wash well) → semantically inaccurate

(e.3) illustrates our concern. If we move the word "好" after the verb and make it become the cohesive complement, it is grammatically correct. Semantically the sentence transfers an idea that is a bit different from the SL. It means "wash the bottles and make them ready for use" while the original meaning should be "wash thoroughly all the bottles to make them clean". This means that if we simply shift the preposed adjunct to a postposed cohesive complement, it does not always resolve the semantic problem. The best choice is either to choose another word which is both semantically and grammatically correct or to employ another kind of syntactic structure rather than the ba-construction. For example, we can choose a more exact word to substitute "好": the word "干净" (clean). For example:

e.4 把所有瓶子洗干净。→ both grammatically and semantically correct
(LWS: ba all bottle thoroughly wash clean)

Or we use other syntactic structures:

e.5 洗干净所有瓶子。(LWS: wash clean all the bottles)

It seems that by selecting the right equivalent and authorizing both syntactic options we could easily resolve this problem. However, if we do so, other obvious problems might reappear. First the French word "bien" is by itself polysemous which contradicts our CL rule of "one word one meaning". Second, it is nevertheless hard to control the use of this highly frequently used word in the SL and we cannot assign many possible equivalents to it. Besides, if we impose on the users by avoiding using this word in many cases, it will make the users very frustrated in looking for alternatives that are not always at hand and thus make our system unfriendly. Furthermore, even if "bien" could be well replaced by "well" in translating it

into English, it will never be the case in Chinese as it is almost impossible to predefine one or even two substitutes that are the equivalents of the SL verbs which happen to be modified by “bien” in French and “好” or “干净” in Chinese. Meanwhile, this approved substitute(s) could also satisfy both the syntactic and semantic restrictions as illustrated above.

In the case of (e.3) it can be regarded as an acceptable translation even if the meaning is not so accurate. For many other cases, it will not be so easily accepted semantically. Besides, some Chinese compound verbs by themselves convey such a meaning and do not need to find an equivalent. We look at the same example by changing the verb and by adding an NP modifier which is also one of our approved syntactic structures:

	French	English
f)	Bien sécher toutes les bouteilles sur la table.	Well dry all the bottles on the table.

Chinese

f.1 把桌子上所有的瓶子晾干。(LWS: ba table on all de[4] bottle dry)

f.2 晾干桌子上所有的瓶子。(LWS: dry table on all de bottle)

Though the word “干” can also be regarded as an adjective, however “晾干” is considered as one word instead of VP. In (f.1) and (f.2), it is not necessary to translate “bien” into any Chinese word though there is no problem in English. If we add “好” either as the adverbial or as the cohesive complement, it will result in a sentence that is only semantically different:

f.3 把桌子上所有的瓶子晾干好。(LWS: ba table on all de bottle dry well; It is good to dry all the bottles on the table.)

f.4 好晾干桌子上所有的瓶子。(LWS: well dry table on all de bottle; <by taking certain measures> it becomes easier/more convenient to dry all the bottles on the table.)

The same situation might also come from the SL. That is to say, a SL word is semantically self-sufficient; however, while translating it into Chinese it is mandatory to have a cohesive/compositive complement which is required by the syntax. For example,

French

g） Premièrement, laver toutes les bouteilles. Sécher les bouteilles et …

English

First wash all the bottles. Dry the bottles and …

Chinese

首先, 把所有的瓶子洗一遍。把瓶子晾干并……

(LWS: first, ba all de bottle wash one time. ba bottle dry and …)

In the above example, the appearance of the cohesive complement (underlined) is mandatory. Therefore, we must add such information restricted by the syntactic-based semantics (which does not exist in the SL) to our system in order that when such sentences are processed, they can be transferred into grammatical ones. To solve these problems, we wish to stress three aspects: 1) extend the lexicon to include more semantic and syntactic information for both verbs and adjectives/adverbs; 2) define syntactic mapping rules to cover all the information needed; 3) establish a syntactic-semantic network to link all this information.

[4] De: a structural particle showing the possessive relationship of the noun, similar to English “of”.

3 Building linguistic models

Up to now we have just given a very narrow and brief description of the basic problems connected to the ba-construction, the adverbial and the cohesive complement. In real texts these syntactic structures exhibit much more complicated variations of which many are not reported in this paper.

As we have already constructed our models for the generation of the ba-construction, we focus on defining other sets of special rules for the construction of the cohesive/compositive complement.

To do this work, we automatically extract the sentences containing the ba-construction and all the verbs that are surrounded either by an adjective or an adverb. We then concentrate on about 30 verbs which can be found in our corpus and which might produce such problems if badly interpreted.

3.1 Lexicalization

In translation, it is well known that the translation between languages often has to deal with the fact of one-to-many and/or many-to-one situations, which is not only the choice of words but also the choice of the syntax. For the human being this is an exciting challenge but also a piece of 'recreation' work. However, this is not the case for MT. Therefore, while searching for possible solutions, we often have to choose a roundabout way in order to minimize the difficulty. One way to solve this problem is to take the task of lexicalization so as to narrow down the semantic gaps between words/phrases from different languages.

In this work, in order to tackle the problems of the semantic incompleteness produced by the syntactic structure, one of our important methods is to have some relatively fixed structures lexicalized. For example, if a SL verb or a SL phrase has to be replaced by a Chinese verb(s) which has to be accompanied by a cohesive/compositive complement, sometimes also by an adverbial, we fix it as one lexical unit to assure the syntax-based semantic completeness. We list these as individual lexical units accordingly, ignoring the fact that in the TL they might be phrases instead of words. It should be noted that in our work the meaning of each word is restricted to no more than one. Unless it is extremely necessary, we do not define more meanings. In the case of necessity, we approve some words to have two or three meanings with strict semantic constraints but it is never allowed to exceed this limit.

Here are a few examples of the lexicalized Chinese equivalents.

French	English	Chinese
mettre en ordre	tidy up	收拾好
mettre … à l'envers	put … upside down	倒置/放
verser entièrement	pour out	倒空
s'allonger (sur le dos)	lie down/lie on one's back	躺平/平躺

Table 3 Examples of lexicalized Chinese equivalents

3.2 Syntax mapping rules

In order to have a set of the syntactic rules that include this information in the TL, the syntactic structures are classified into different categories which might be further divided into several sub-categories accordingly so as to include the possible reconstruction information of the approved syntax. This approach is again supported by set theory. Here are some simple examples of part of the approved syntactic structures. To make it easier to read, we only choose a few syntactically very simple examples to show the basic construction of some types of the approved phrasal structures.

Here are some of our approved syntactic structures in relation to the Ba-construction.

NP1 + ba +NP2 + VP, in which the VP could be:

1) VP → (PP) + V + PP, e.g.

French: Placer le flacon sur la table. English: Put the flask on the table.

Chinese: 把烧瓶放到桌子上。(LWS: ba flask put to table shang- Loc[5])

2) VP → V + CC[6], e.g.
Fr: Laver bien les bouteilles. En: Wash well the bottles.
Ch: 把瓶子洗干净。(LWS: ba bottle wash clean)

3) VP → (AdP) + V + CC, e.g.
Fr: Vider complètement l'eau du seau. En: Pour out the water from the bucket.
Ch: 把桶里的水(彻底)倒尽。(LWS: <completely> ba bucket Loc <inside> de water pour <finish>)

4) VP → V + NP + CC (consisted by verbal classifier), e.g.
Fr: Répéter la première procédure. En: Repeat the first procedure.
Ch: 把第一个步骤重复一遍。 (LWS: ba first procedure repeat one bian-classifier)

For the verbs that could not be used in the ba-construction, we also define the approved structures. Here we take only one of them as an example.
(NP) + VP, in which VP could be:

5) VP → AdP + V + NP + (PP), e.g.
Fr: Observer attentivement le changement de couleur de l'eau.
En: Observe carefully the change in the color of the water.
Ch: 仔细观察水的颜色变化。(LWS: carefully observe water de color change)

Of course there are still many other types of structures that could function either as an adverbial or as a cohesive/compositive complement as for instance, the verbal complement which consists of particles ("了", "着") and/or set expressions.

3.3 Semantic tags and equivalent candidates in the lexicon

In order that the system can make some wise and exact decisions while translating French sentences into Chinese ones, we also add more information to the lexical units in the lexicon. For example, we have already defined the syntactic information related to verbs to indicate what kind of syntactic structure a verb can have. We then list one or two candidate words that can function either as the adverbial or as the cohesive complement. Here are two simple examples showing how these pieces of information are linked through different linguistic databases.

Lexical entry	SynInfo	SemInfo	AdjInfo
Fr: observer Ch: 观察	±ba	<human behavior>	Attentivement; bien 认真+AdP/−CC ; 仔细+AdP/+CC
Fr: poser Ch: 放	+ba	<human behavior>	Doucement 轻+AdP/ −CC

Table 4 Examples of the lexical entry

In this table we only list the information to show whether a Chinese verb can be used in the ba-construction or not. "±ba" means that the verb can be used both in the ba-construction and the structure without it. "+ba" means that the verb can only be used in the ba-construction. Therefore if a verb is

[5] Loc: a noun of locality signifying the location, often used together with a preposition, here "shang" means "on the surface of".
[6] CC: short for the cohesive complement

marked as "−ba", it means it is not allowed to appear in the ba-construction. Similarly, while "仔细 +AdP/+CC" means that this adjective can take both the adverbial position and the cohesive complement position, "认真+AdP/−CC " means that "认真" can only take the adverbial position in our presently approved syntax. This information together with information stored in the grammar rule base assures that our system makes the right choices while generating Chinese sentences.

4 Evaluation

The evaluation of the work is done by using our own system. The first reason for such a choice is that the ba-construction is often badly translated using other commercialized or online free MT systems. In fact in most cases, these systems seldom succeed in generating a grammatical Chinese sentence containing the ba-construction. As a result, we cannot collect useful information for the improvement of our own system. Secondly, even if some sentences are well translated, it is still hard to set any standard for the evaluation purpose. Besides, though the linguistic phenomena we present in this paper are very important for a better translation, these are still not well implemented by most MT systems. Therefore, our only choice is to adapt our own system to take over the task.

We set three standards for the evaluation: good, acceptable and bad for structure, word choice and meaning transfer. We select only twenty sentences for the evaluation, of which twelve have to be generated into Chinese by employing the ba-construction while the other eight sentences must be generated into sentences that could not use the ba-construction. Our result is nevertheless satisfactory with 98% accuracy compared with the first translation without these new pieces of linguistic information. In the first translation, almost all the sentences that need a cohesive complement are not correctly translated with an accuracy of less than 60%. However, we must stress here is that within a domain-oriented task with well controlled syntax and lexical meanings, such an accuracy is not surprising.

5 Conclusion

In this paper, we have explained part of the work that we have done for the MT of well controlled sentences from French to Chinese by illustrating the problems produced by certain language specific particularities. Based on our analysis of these linguistic phenomena, we have managed to build linguistic models to solve the problems in order to improve our MT quality.

References

Cardey, Sylviane. 2009. *Controlled Languages for More Reliable Human Communication in Safety Critical Domains.* Proceedings of the 11th International Symposium on Social Communication, Santiago de Cuba, Cuba, January 19-23, 2009, pp. 330-335.

Cardey, Sylviane. Greenfield, Peter. Wu, Xiaohong. 2004. *Designing a Controlled Language for the Machine Translation of Medical Protocols: the Case of English to Chinese.* Proceedings of the AMTA 2004, LNAI/3265, Springer-Verlag, pp. 37-47.

Milićević, Jasmina. 2006. *A Short Guide to the Meaning-Text Linguistic Theory.* Journal of Koralex, Vol. 8, pp. 187-233.

Wu, Xiaohong. 2005. *Controlled Language – A Useful Technique to Facilitate Machine Translation of Technical Documents,* Lingvisticæ Investigationes 28:1, 2005. John Benjamins Publishing Company. pp. 123-131.

Wu, Xiaohong. Cardey, Sylviane. Greenfield, Peter. 2006. *Some Problems of Prepositional Phrases in Machine Translation*. Proceedings of the 5th International Conference on Natural Language Processing, FinTAL, Turku, Finland, 2006, Springer-Verlag – LNAI 4139, ISBN 3-540-37334-9, pp. 593-603.

Ye, Zheng. Zhan, Weidong. Zhou, Xiaolin. 2007. *The Semantic Processing of Syntactic Structure in Sentence Comprehension: An ERP Study,* Brain Research, Volume 1142, 20 April, 2007, pp. 135-145.

Zhan, Weidong. 2004. 广义配价模式与汉语“把”字句的句法语义规则. 语言学论丛 No 29, pp. 334-368, Commercial Press 2004.

Types of Paraphrase Rules in Practice. German Paraphrases of a Russian Text

Robert Zangenfeind
Centrum für Informations- und Sprachverarbeitung (CIS)
Ludwig-Maximilians-Universität
Oettingenstraße 67
80538 München
Germany
R.Zangenfeind@lmu.de

Abstract

Paraphrases can be described at various levels of utterance representation in the Meaning-Text Model (MTM). We analyze different kinds of paraphrases that are found in a parallel corpus of German translations of a Russian novel and assign them to the different representation levels. Doing so, some rules of the deep syntactic paraphrasing system of MTM are modified. For paraphrases that need a deeper analysis to be recognized as such a division into two groups is proposed: paraphrasings that can be described at the semantic level and those that can be described at the transition between the semantic and the deep syntactic level. Some difficulties in the formal recognition of paraphrases will be discussed and the frequency of different types of paraphrase rules will be evaluated.

1 Introduction

Paraphrasing and synonymy of natural language texts can be observed at every level of utterance representation in the MTM. There is synonymy which can be described at the level of morphological representation,[1] of surface syntactic representation, of deep syntactic and of semantic representation, cf. (Apresjan & Cinman, 2002:104ff.) and (Milićević, 2007a:128–138).

In this paper we will examine to what extent paraphrase rules at these different levels are applied in practice, i.e. what kind of paraphrases are to be found when analyzing real texts. As a basis for this research we need a parallel corpus of texts having the same meaning. Such a corpus can be found e.g. in different translations of an original text which are especially available of classical novels. Thus, the original text represents a given meaning which is then expressed in different translations that represent paraphrases of each other. (Wirth, 1996) provides such a corpus by 22 German translations of the first ten sentences of Lev N. Tolstoj's novel Anna Karenina. The essential basis for what follows was a comparison of the 21 main predicates of this opening part of the novel, i.e. predicative nouns, adjectives and adverbs in support verb constructions, full verbs, and idioms. Some of the paraphrases of these predicates will be presented here. Besides that, paraphrases of some attributes and actants of these predicates and some pragmatic aspects will be considered. All linguistic examples are taken from (Wirth, 1996) unless noted otherwise.

[1] We will not distinguish between the deep and the surface morphological level in this paper because it is not necessary for our purpose.

2 Paraphrases at different levels of representation

A very important part of rules for paraphrasing natural language texts is described by means of lexical functions in the paraphrasing system which operates at the deep syntactic level of the MTM, cf. (Mel'čuk, 1974:149–176; Mel'čuk, 1992; Apresjan, 1974:316–345; Apresjan & Cinman 2002) and section 3 in this paper. Paraphrasing at the semantic level within the MTM has only recently been described, cf. (Milićević, 2005, 2007a, 2007b) and section 4 in this paper. A different kind of paraphrase – which has not been studied so far in detail – cannot be described within one single level but should be described at the transition between the semantic and the deep syntactic level, cf. section 5. Paraphrases at levels of representation that are closer to surface are briefly discussed in section 6. These include paraphrasings of surface syntactic representations which can be described e.g. by means of different realizations of connecting actants with predicates to their key word. Morphological paraphrases are realized by alternative spellings of a certain lexeme.

3 Paraphrases of predicates at the deep syntactic level

3.1 Exact paraphrases of predicates

First of all, we want to examine some paraphrases that can be described by means of lexical functions at the deep syntactic level.

The lexical rule 24 of (Mel'čuk 1992:39) which describes the paraphrase of a full verb into a support verb construction with a predicative adjective, $C_{0(V)} \Leftrightarrow A_1(C_0) \leftarrow\!\!\text{II}\!\!- \text{Oper}_1(A_1(C_0))$, is applied e.g. in the paraphrase of the following predicate:

(1) a. alle glücklichen Familien *ähneln* [C_0] einander 'all happy families resemble each other'
⇔
b. alle glücklichen Familien *sind* $\text{Oper}_1(A_1(C_0))$ einander *ähnlich* [$A_1(C_0)$] 'all happy families are similar to each other'

This rule and further rules for support verb constructions with predicative nouns as well as rules for synonyms, antonyms, and conversives which are applied quite often in our corpus, are described completely within the paraphrasing system of MTM.

3.2 Approximate paraphrases of predicates

The lexical rule 1 of (Mel'čuk 1992:37), $C_0 \Leftrightarrow \text{Syn}(C_0)$, is applied very often with the modification that not a genuine synonym but a partial synonym is used in the paraphrase, as in (2):

(2) a. alle glücklichen Familien *ähneln* [C_0] einander 'all happy families resemble each other'
⇔
b. alle glücklichen Familien *gleichen* [$\text{Syn}_{\subset}(C_0)$] einander 'all happy families equal each other'

The formal notation of this rule could be rule 1':

Rule 1' $C_0 \cong \text{Syn}_{\subset}(C_0) \cong \text{Syn}_{\supset}(C_0) \cong \text{Syn}_{\cap}(C_0)$

This rule can be used, of course, not only for predicates but also for actants and attributes.

There are some rules which are part of the semantic implications in Mel'čuk's (1992) system, namely rules 53 and 54, IncepX => X and ContX => X. When the application of rule 53 is combined with a change of tense and rule 54 is varied into rule 54' (cf. below) they describe a semantic implication in both directions and thus can form a class of approximate paraphrases. This can be illustrated in (3) and (4) respectively:

(3) a. der Mann *war* [C_0] nicht zu Hause 'the husband was not at home'
⇔

b. der Mann *war* nicht nach Hause *gekommen* [Incep(C_0)] ‘the husband had not come home’

The German verb *sein* ‘be’ in its Imperfekt form *war* ‘was’ can be paraphrased very well into the verb *kommen* ‘come’ which is an Incep(*sein*) when used in its Plusquamperfekt form[2] *war gekommen* ‘had come’ and the other way round. So, a variation of rule 53 could be rule 53’:

Rule 53’ $C_{0(V,Imperfekt)} \cong Incep(C_0)_{Plusquamperfekt}$[3]

The rule 54 could be varied into rule 54’ by adding a temporal adverb on both sides of the rule because the presence of a temporal adverb in both phrases produces fairly good paraphrases in both directions, like in (4).

Rule 54’ $C_{0(V)}$ [+ Adv_{temp}] ≅ $Cont(C_0)$ [+ Adv_{temp}]

(4) a. Ivan *war* [C_0] drei Tage in Odessa ‘Ivan was in Odessa for three days’
⇔
b. Ivan *blieb* [$Cont(C_0)$] drei Tage in Odessa ‘Ivan stayed in Odessa for three days’[4]

In a similar way these approximate rules are applied in our corpus to different kinds of support verb constructions. Examples (5) and (6) may serve as an illustration for this:

(5) a. die englische Gouvernante *zankte sich* [C_0] ... ‘the English governess quarreled ...’
⇔
b. die Erzieherin war *in Zank* [$S_0(C_0)$] *geraten* [$IncepOper_1(S_0(C_0))$] ... ‘the governess had got into a quarrel ...’[5]

(6) a. Es war nun schon der dritte Tag, dass diese Situation [C_0] *bestand* [$Func_0(C_0)$] ‘it was the third day that this situation existed’
⇔
b. Es war schon der dritte Tag, dass diese gespannte Situation [C_0] *fortdauerte* [$ContFunc_0(C_0)$] ‘it was the third day that this tense situation continued’

A lot more examples of this kind of deep syntactic paraphrasing are analyzed in (Zangenfeind, 2009).

3.3 Syntactic paraphrase of an attributive adjective into a predicate one

There are also some simply syntactic paraphrases that don’t imply any lexical functions (similar to active – passive transformations) like e.g. in (7) where an attributive adjective is paraphrased into a predicative one in a relative clause:

(7) a. alle *glücklichen* [C_0] Familien ‘all happy families’
⇔
b. alle Familien, *die glücklich* [C_0] *sind* [$Copul(C_0)$] ‘all families that are happy’

---

[2] In some contexts the Perfekt form also produces quite acceptable paraphrases.
[3] Two points about this rule should be specified: firstly, we are aware that the rule in this form is not universal – in another language this rule must be adopted according to its tense system. In Russian e.g. an appropriate change of tense can be realized by a change from the imperfective aspect to the perfective aspect like in: он *спал* [C_0] ≅ он *заснул* [$Incep(C_0)$] ‘he slept ≅ he had fallen asleep’. More research is needed to formulate a more general rule. Nevertheless, even if we don’t have a universal rule now, we have to take into account that a considerable number of paraphrases of this type are found in real texts. Secondly, in practice this rule is sometimes applied even without change of tense which, of course, makes the paraphrase less exact and more approximate.
[4] This example is not taken from Anna Karenina but contributed by the author.
[5] In this paraphrase a synonym for this 1st actant is used, *Erzieherin* = Syn(*Gouvernante*) ‘governess’, and the attribute *englisch* ‘English’ is omitted, cf. about this problem below.

The deep syntactic rule for this paraphrase is a general one and can be noted as in figure 1:

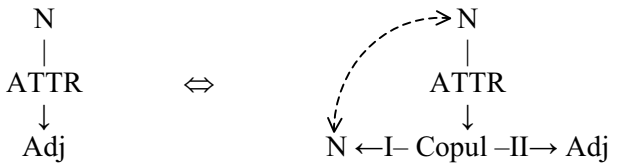


Figure 1. Syntactic rule for paraphrasing an attributive adjective into a predicative one.

It would be possible as well to describe the paraphrase in (7) as a transition rule between the semantic representation and the deep syntactic representation, like some others, as will be shown below.

3.4 Paraphrases of idioms

A lot of paraphrases by means of an idiom can be described at the deep syntactic level because they correspond to a single node of the deep syntactic structure. This is possible e.g. with the paraphrase in (8):

(8) a. die Frau hatte *erfahren*, dass … 'the wife had learned that …'
⇔
b. die Frau hatte *Wind bekommen* von ... 'the wife had got wind of …'

The idiom *Wind bekommen* 'get wind' in (8b) corresponds to the single node *erfahren* 'learn' which can be noted in the dictionary as Syn(*Wind bekommen*) = *erfahren*. However, this kind of notation is not possible for all idioms. An example to illustrate this is given in (9):

(9) bei jemandem ist Hopfen und Malz verloren '*lit*. hops and malt are lost on somebody'[6]

This idiom roughly means 'it is not worth the trouble trying to help or reform somebody because it seems useless'. This meaning cannot be expressed in German by a single lexeme, and thus cannot be noted comfortably in terms of the lexical function Syn. As a consequence, paraphrases like this must be described at the transition between the semantic and the deep syntactic level, cf. below.

This raises the question whether the paraphrasing of idioms should formally be described in two different ways. Of course, to avoid this, all paraphrasings by an idiom in principle could be described at the transition between the semantic and the deep syntactic level – which would be awkward in all cases where it is possible to note them in terms of a synonym (which is possible in most cases). But that is a problem that cannot be dealt with sufficiently here.

4 Paraphrases at the semantic level

4.1 Approximate paraphrases

A lot of paraphrases in the translations of Anna Karenina are rather complicated and the only reasonable way to describe them could be with rules at the semantic level because they have different semantic representations. These rules are complicated in themselves, cf. some rules in (Milićević, 2007a, 2007b) but what makes them additionally complicated, is the fact that in practice these paraphrases are sometimes only very approximate paraphrases, like the following:

(10) a. alle glücklichen Familien *gleichen* [C_0] einander 'all happy families equal each other'
⇔

[6] Example by the author.

b. *die äußeren Formen*, in denen das sogenannte glückliche Familienleben sich abzuspielen pflegt, *sind* überall *die gleichen* 'the external forms in which the so called happy family life usually takes place are the same everywhere'

To describe paraphrases like these in a formal way and recognize them automatically seems hardly possible because, other than in paraphrases which Milićević (2007a) treats as propositional quasi-equivalences, the semantic relation between paraphrases like in (10) is quite irregular. Of course, one may ask whether (10b) is really a well-formed paraphrase of (10a) (which is a good translation of the Russian original phrase) but, nevertheless, it is a paraphrase that is found in a real text.

4.2 Paraphrases concerning pragmatics

In (11b) an introductory expression attracts the attention of the reader in a pragmatic way without giving additional meaning to the sentence concerning the extralinguistic situation. Nevertheless, the two phrases of (11) have different semantic representations because of the speaker's attitude that is communicated in (11b) and must be described in the modal frame[7] which is roughly 'P, and the speaker thinks that P is important':

(11) a. die Frau des Hauses hatte erklärt ... 'the hostess had announced ...'
⇔
b. *die Sache war die, dass* die Frau des Hauses erklärt hatte ... 'the point was that the hostess had announced ...'

In the phrases of (12) a change of communicative organization takes place by means of a change of word order. This does not affect the semantic structure of the phrases but only the distribution of the semantic theme and rheme, cf. (Mel'čuk, 2001:61f.):

(12) a. im Hause der Oblonskijs ging alles drunter und drüber 'at Oblonskijs' house everything was at sixes and sevens'
⇔
b. alles ging drunter und drüber im Hause der Oblonskijs 'everything was at sixes and sevens at Oblonskijs' house'

Another change of pragmatics occurs in the paraphrase of (13) by the addition of the particle *mehr* 'more' which must be described again in the modal frame of the semantic representation:

(13) a. der Mann war seit drei Tagen nicht nach Hause gekommen 'the husband had not come home for three days'
⇔
b. der Mann war seit drei Tagen nicht *mehr* nach Hause gekommen 'the husband had not come home again for three days'

The phrase (13a) is more neutral than (13b) which includes a presupposition of the speaker who thinks that three days is a lot. The propositional meaning that is described in both phrases, i.e. 'the man was not at home for three days', is not affected by this.

5 Paraphrases of predicates at the transition between the semantic and the deep syntactic level

There are some paraphrases which cannot be described sufficiently at one single level of representation. The reason for this is that on the one hand they have different deep syntactic representations but (more or

[7] Cf. e.g. Apresjan (1974:67).

less) the same semantic representation,[8] which means that no paraphrase at the semantic level is necessarily required. On the other hand their different deep syntactic representations cannot be paraphrased into each other by means of lexical functions because they do not belong to the same so called basic deep syntactic structure (базовая ГСС, cf. Mel'čuk, 1974:177).[9] The only way how phrases of this kind can be recognized as paraphrases is to analyze them down to their semantic representation. This kind of paraphrases has not yet been discussed in detail, as far as we know.

The two phrases in (14) also have more or less the same semantic representation (maybe with a hint to slightly different stylistic specifications in the rhetorical structure). Their paraphrasing implies a substitution of a predicate which at first glance could be expressed by a simple lexical rule with a lexical function, namely $C_{0(V)} \Leftrightarrow S_0(C_0)$. Nevertheless, this paraphrase is more complex and should be formalized at the transition between the semantic and the deep syntactic level when the nodes of the semantic net are replaced by deep syntactic units. This is because it is rather complicated to describe the relation of the semantic actants of the different predicates involved in the paraphrase:

(14) a. die englische Erzieherin hatte einer Freundin einen Brief geschrieben, *in dem sie dieselbe bat* [C_0], ihr eine neue Stelle zu besorgen '*lit.* the English governess had written a letter to a friend in which she asked her to get her a new job'

⇔

b. die englische Gouvernante hatte einer Freundin einen Brief geschrieben *mit der Bitte* [$S_0(C_0)$], ihr eine neue Stellung zu verschaffen '*lit.* the English governess had written a letter to a friend with the request to get her a new job'

The 1st semantic actant of *Brief* 'letter' (*die englische Erzieherin/Gouvernante* 'the English governess') is realized in both phrases as the syntactic actant I of *schreiben* 'write' which serves as $Oper_1$(*Brief*). This actant is also the 1st semantic actant of *bitten* 'ask' in (14a) and of *Bitte* 'request' in (14b). But, while this is explicated by the pronoun *sie* 'she' in (14a), it is only implicitly known in (14b). Similarly, the 2nd semantic actant of the request (*Freundin* 'friend') is explicitly realized in (14a) by the pronoun *dieselbe* 'the same' and only implicitly known in (14b). These relations are clear in the semantic representation, and the way they can be realized or omitted in the text should be formally noted in a rule of the semantic component of the MTM, i.e. at the transition from the semantic to the deep syntactic representation. This rule must describe the following fact: When some kind of information (like e.g. a request, a promise, news) is transported by a "bearer/medium of the information" (which can be e.g. a letter, a telegram or an agreement) the 1st actant of the information is identical with the 1st actant of the medium of the information. The same applies to the counter agent of the information which is identical in such cases with the counter agent of the medium of the information.

The paraphrase in (15) shows two phrases with the same request as in (14); in (15a) the request is explicitly expressed by means of the full verb *bitten* 'ask', similarly to (14a), in (15b) it is expressed by the modal verb *mögen* 'may' in a slightly different syntactic construction:

(15) a. die englische Erzieherin hatte einer Freundin geschrieben, sie *bitte* [C_0] sie, sich nach einer anderen Stellung für sie umzutun '*lit.* the English governess had written to a friend she would ask her to look for a new job for her'

⇔

b. die englische Erzieherin hatte an eine Freundin geschrieben, sie *möchte* sich nach einer anderen Stelle für sie umsehen '*lit.* the English governess had written to a friend she [the friend] may look for a new job for her'

The use of *mögen* 'may' in this context of a request should also be described as a rule of the semantic component of the MTM because a paraphrase of *bitten* 'ask' by *mögen* 'may' is not possible in every con-

[8] In this respect they are similar to conversives which have the same semantic structure but different deep syntactic structures (and different communicative structures), cf. (Mel'čuk, 2001:120f.).

[9] In this respect they are different from conversives which do belong to the same basic deep syntactic structure.

text. In a phrase like (16) where the 3rd actant of the request is not realized by a verbal expression it would produce an ungrammatical sentence:

(16) a. die Erzieherin hatte einer Freundin geschrieben, sie *bitte* den Hausherrn um eine Stelle für Ivan '*lit.* the governess had written to a friend she would ask the landlord for a job for Ivan'[10]
⇔
b. *die Erzieherin hatte einer Freundin geschrieben, der Hausherr *möchte* eine Stelle für Ivan '**lit.* the governess had written to a friend the landlord may a job for Ivan'

6 Paraphrases at the surface syntactic level and at the morphological level

In the government pattern of predicates different forms of connecting actants to a keyword are described. This allows the formal description of paraphrases at the surface syntactic level, like e.g. in (17). Unlike the paraphrases above this is not a paraphrase of the predicate itself:

(17) a. die englische Erzieherin hatte *einer Freundin* geschrieben 'the English governess had written a friend'
⇔
b. die englische Erzieherin hatte *an eine Freundin* geschrieben 'the English governess had written to a friend'

Here the verb *schreiben* 'write' is used as a full verb. In its government pattern the following two possibilities for a connection of its 3rd actant, the counter agent, are described: a noun in the dative case and a noun in the accusative case via the preposition *an* 'to' (or formally: D3.1 N_{dat} and D3.2 *an* N_{acc}).[11]

The change of word order in (18) is a paraphrase without change of communicative organization at the semantic level, other than in (12). The two phrases of (18) have identical surface syntactic structures but different morphological structures and thus can be described formally at the transition between the surface syntactic and the morphological level.[12]

(18) a. die Hausfrau hatte erklärt, sie könne nicht länger *mit ihm unter einem Dache* wohnen 'the hostess had announced that she could not live any longer with him under one roof'
⇔
b. die Frau hatte erklärt, sie könne nicht mehr *unter einem Dache mit ihm* wohnen 'the hostess had announced that she could not live any longer under one roof with him'

Paraphrases at the morphological level concern e.g. alternative inflections and spellings of lexemes which must be noted in the morphological zone of the dictionary entry; an example of this is shown in (19) where the lexeme *Dach* 'roof' is realized in two different dative forms:

(19) a. unter einem *Dach* 'under one roof'
⇔
b. unter einem *Dache* 'under one roof'

In principle, paraphrases like these are possible, of course, with predicates as well as with actants and attributes like the one in (19). They can be identified without difficulty because of their formal description in the dictionary.

---

[10] Example by the author.
[11] More types of paraphrases at the surface syntactic level are described in (Milićević, 2007:135f.). All these paraphrases don't concern predicates directly.
[12] (Milićević, 2007:136f.) describes a similar paraphrase with a change of linear word order at the deep morphological level. Nevertheless, as long as there are no explicit rules for paraphrases by means of word order at the deep morphological level, an analysis to the surface syntactic level is necessary to recognize paraphrases like these because the information for different forms of linearization of an utterance is available only in the surface syntactic representation.

A special problem in this context, however, can be caused by the different spellings of transcribed Russian names and other expressions because they often do not have dictionary entries. The name of the family in Tolstoj's novel e.g. has three different spellings in the translations: *Oblonskij*, *Oblonski* and *Oblonsky*.

7 Two difficulties in the recognition of paraphrases

When examining the German paraphrases of Anna Karenina, two difficulties are met quite often: Firstly, in a lot of paraphrases that can be described by lexical rules of the deep syntactic paraphrasing system of the MTM two or even more lexical rules are applied at the same time for one single predicate. Secondly, as already has been stated above, a lot of paraphrasings in practice are only very approximate paraphrases at the semantic level or they are based on only very vague lexical synonyms, some of which are not to be found in the dictionary.

An example for a simple combination of two rules is shown in (20):

(20) a. alle *fühlten* ... 'everybody felt'
⇔
b. alle *hatten das Empfinden* ... 'everybody had the emotion'

Here the full verb *fühlen* 'feel' is paraphrased by a support verb construction with an $Oper_1$ and a predicative noun. At the same time the predicative noun S_0(*fühlen* 'feel') – which is *Gefühl* 'feeling' – is paraphrased by Syn(*Gefühl* 'feeling') which is *Empfinden* 'emotion'.

A quite complex combination of rules in just one phrase that is part of a relative clause in reported speech is shown in (21):

(21) a. ... dass die in einer Herberge zufällig sich begegnenden Gäste [E1] *mehr* [$C_{0\text{-}2,kompar}$] miteinander *verbunden* [$A_1(C_{0\text{-}1})$] *seien* [$Oper_1(A_1(S_{0\text{-}1}))$] als sie [die Familienmitglieder = E2] '... that the guests that meet accidentally in a hostel are connected more with each other than they [the members of the family] are'
⇔
b. ... dass *kaum* noch *so viel* [$Anti(C_{0\text{-}2,kompar})$] *Gemeinsames* [= $Syn_\cap(C_{0\text{-}1})$] zwischen ihnen [E2] *war* [$Func_0(Syn(S_{0\text{-}1}))$], wie etwa zwischen den zufälligen Gästen [E1] eines Hotels '... that there was hardly as much common ground between them anymore as, for instance, between the accidental guests of a hotel'

There are two predicates that are essential for a consideration of this paraphrase. In (21a) one predicate is realized by the lexeme *verbunden* 'connected' [= $A_1(C_{0\text{-}1})$], an adjective that describes the property of the first actant of the situation C_0 which could be represented by *Verbindung* 'connection'. This adjective is part of a support verb construction with the $Oper_1$ *sein* 'be' in its subjunctive form *seien*. The other predicate is the lexeme *mehr* 'more' [= $C_{0\text{-}2,kompar}$], a comparative adverb. The actants of the comparison are *Gäste* 'guests' [= E1] and *sie* 'they' [= E2] (a pronoun that represents the members of the family).

In (21b) the first predicate is realized by a partial synonym of C_0, namely *Gemeinsames* 'common ground' [= $Syn_\cap(C_{0\text{-}1})$] which is used in a support verb construction with $Func_0$. The second predicate is realized now as an antonym by means of *kaum so viel* 'hardly as much as' [$Anti(C_{0\text{-}2,kompar})$], i.e. 'less' which entails an exchange of its actants *Gäste* 'guests' and *sie* 'they' in its dative form *ihnen*.

So, the rules for a formal description of the paraphrase of these two predicates are a combination of rules 20 and 24 from (Mel'čuk, 1992), $A_1(C_0)$ ←II– $Oper_1(A_1(C_0))$ ⇔ $S_0(C_0)$ ←I– $Func_0(S_0(C_0))$, the modified rule 1', $C_0 \cong Syn_\cap(C_0)$ (cf. above), and a rule from (Apresjan & Cinman, 2002), E1 + $C_{0,kompar}$ + E2 ⇔ E2 + $Anti3(C_{0,kompar})$ + E1. Written in one rule this would be (shown here without syntactic relations): $A_1(C_{0\text{-}1})$ + $Oper_1(A_1(C_{0\text{-}1}))$ + E1 + $C_{0\text{-}2,kompar}$ + E2 ⇔ $Syn_\cap(C_{0\text{-}1})$ + $Func_0(Syn_\cap(C_{0\text{-}1}))$ + E2 + $Anti(C_{0\text{-}2})$ + E1.

But the paraphrase of (21) is even more complex, and here we come to the second difficulty of recognizing paraphrases in practice. The first actant, E1, is specified in (21a) by the expression *die in einer Herberge zufällig sich begegnenden [Gäste]* 'the guests that meet accidentally in a hostel' which is

considerably different from the specification in (21b): *die zufälligen [Gäste] eines Hotels* ‘the accidental guests of a hotel’. The lexeme *Herberge* ‘hostel’ does not really mean the same as *Hotel* ‘hotel’ and the attribute *zufällig sich begegnend* ‘that meet accidentally’ is more precise than simply *zufällig* ‘accidental’. Furthermore, in (21b) there is a modal frame by the lexeme *noch* ‘anymore’ and the parenthesis *etwa* ‘for instance’ which make the paraphrase still more approximate.

In other paraphrases some actants are even more different from each other than they are in (21). So, in some translations instead of *Gäste* ‘guests’ the very distant synonym *Leute* ‘people’ is used. And there are predicates which are used in paraphrases but are only very vague synonyms, like e.g.:

(22) a. *Sinn* ‘sense’
⇔
b. *höherer Gedanke* ‘higher idea’
⇔
c. *innerer Zusammenhang* ‘internal connection’
⇔
d. *Boden* ‘ground’

It is fairly obvious that paraphrases like these are really hard to recognize automatically.

In practice, as we have seen in (21), the two discussed difficulties are met quite often at the same time in one single phrase. In fact, it is hard to find phrases in which just one single rule is applied. This makes the practice of recognizing paraphrases in real texts even more complicated.

8 Evaluation of the frequency of rules

When considering the 21 predicates of the first ten sentences in the German translations of Tolstoj’s novel we find about 46% are no paraphrases of each other but use identical predicates – these are the direct translations of the predicates in the Russian original text. About 19% of the predicates are paraphrases by means of synonyms, including idioms (5%), about 14% of the paraphrases can be described by exact rules for support verb constructions, 7% of the paraphrases must be described at the semantic level, 5% at the transition between the semantic and the deep syntactic level, for a little more than 3% of the paraphrases approximate rules for support verb constructions are applied, for a little less than 3% approximate rules for full verbs, and 2% are paraphrasings by an antonym or conversive.

Leaving aside direct translations and taking into consideration only paraphrases of these, this leads to the following distribution: Almost 79% of all paraphrases which are found in the 22 translations of Anna Karenina can be described by lexical rules of the deep syntactic paraphrasing system of the MTM. Most interesting is that a big part of this, i.e. more than 32% (including approximate paraphrases) of all rules applied, are rules for support verb constructions – that is almost as much as the 36% paraphrases by synonyms. 12% of the predicates are paraphrased by rules that are applied at the semantic level and 9% can be described by rules applied at the transition between this level and the deep syntactic representation.

These figures show very impressively the importance of the deep syntactic paraphrasing system of the MTM, especially of the rules for support verb constructions. Yet, the considerable numbers of semantic paraphrasings and of those that are described by transition rules show that it is important too, on the other hand, to develop further these parts of the MTM.

Acknowledgements

Many thanks to Franz Guenthner for the discussions during the preparation of this paper and to the two anonymous reviewers for their comments on the pre-final version of this paper.

References

Mel’čuk, Igor A. 1992. Paraphrase et lexique. In: Igor A. Mel’čuk et al., *Dictionnaire explicatif et combinatoire du français contemporain: Recherches lexico-sémantiques, vol. III*. Montréal: Presses de l’Univ. de Montréal, 9–58.

Mel’čuk, Igor A. 2001. *Communicative Organization in Natural Language: The Semantic-Communicative Structure of Sentences*. Amsterdam: John Benjamins.

Milićević, Jasmina 2005. Semantic Paraphrasing in Meaning-Text Linguistic Models. In: Юрий Д. Апресян (ред.), *Восток – Запад*. Москва: Языки славянской культуры, 316–329.

Milićević, Jasmina 2007a. *La paraphrase. Modélisation de la paraphrase langagière*. Bern: Peter Lang.

Milićević, Jasmina 2007b. Semantic Equivalence Rules in Meaning-Text Paraphrasing. In: Leo Wanner (ed.), *Selected Lexical and Grammatical Issues in the Meaning–Text Theory*. Amsterdam: John Benjamins, 267–296.

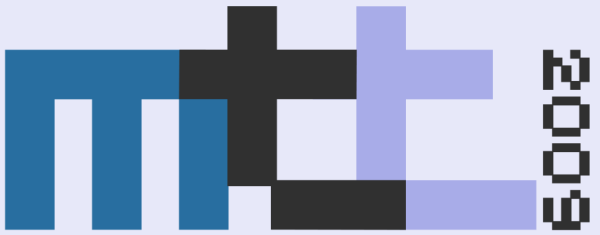
Wirth, Dieter 1996. *Paraphrase und Übersetzung in einem Inhalt ↔ Text-Modell*. Tübingen: Niemeyer.

Zangenfeind, Robert 2009. *Das Bedeutung-Text-Modell und dessen Paraphrasierungssystem unter besonderer Berücksichtigung von Stützverbgefügen*. [To appear]

Апресян, Юрий Д. 1974. *Лексическая семантика. Синонимические средства языка*. Москва: Наука.

Апресян, Юрий Д. & Леонид Л. Цинман 2002. Формальная модель перифразирования предложений для систем переработки текстов на естественных языках. In: *Русский язык в научном освещении*, № 2 (4), 102–146.

Мельчук, Игорь А. 1974. *Опыт теории лингвистических моделей «Смысл ⇔ Текст». Семантика, синтаксис.* Москва: Наука.



meaning text theory